École doctorale
Sociétés, politique, santé publique

Apprentissage supervisé pour données massives multi-blocs incomplètes

Hadrien Lorenzo[⊙], Jérôme Saracco[⋆], Rodolphe Thiébaut[⊙]

⊙ Univ. Bordeaux, Inria Bordeaux Sud-Ouest-France (*BSO*), Inserm U1219, Bordeaux Population Health Research Center, SISTM team
⋆ CQFD Inria BSO, CNRS (UMR5251), Bordeaux INP
E-mail : hadrien.lorenzo@u-bordeaux.fr

Mots-clés : Données multi-blocs, Données hétérogènes, Données manquantes, Apprentissage supervisé, PLS, Seuillage doux, Sélection de variables, Multi-Omique

Introduction

Les récentes innovations techniques ont permis la production de données massives en biologie, comme les données omiques par exemple (génomique, transcriptomique, protéomique, ...). Ces données peuvent être manquantes pour des raisons techniques (mauvaise qualité de l'extraction de l'ARN à un temps de prélèvement par exemple). Or, un transcriptome manquant signifie la perte de l'expression de milliers de gènes. Nous avons rencontré ce problème pour l'analyse de données de vaccination Ebola rVSV [5] où plus de 30% de l'information était manquante. Avec ces données de vaccination Ebola, la question est de prédire la réponse anticorps des patients, c'est donc un problème supervisé.

Objectifs de ddsPLS

mddsPLS [4] (*multi-block data driven sparse Partial Least Squares*) :

- Gérer les données manquantes,
- Données multi-blocs hétérogènes de grande dimension,
- Sélection de variables,
- Régression ou Classification.

Validation de la méthode par comparaison à des méthodes classiques :

- Simulations numériques,
- Applications à plusieurs jeux de données réelles.

Packages **R** et **Python** [a].

Méthodologie

T types de données décrivant n individus via p_t, $\forall t \in [\![1, T]\!]$ variables. Exemples :

- **ARN**. Expression de l'ARN séquencé.
- **Clinique**. Valeurs phénotypiques (âge, poids, taille, état clinique général,...).
- **SNP**. Polymorphismes nucléaires simples, homozygotie.
- **Protéines**. Activité protéique, quantités ou concentrations.

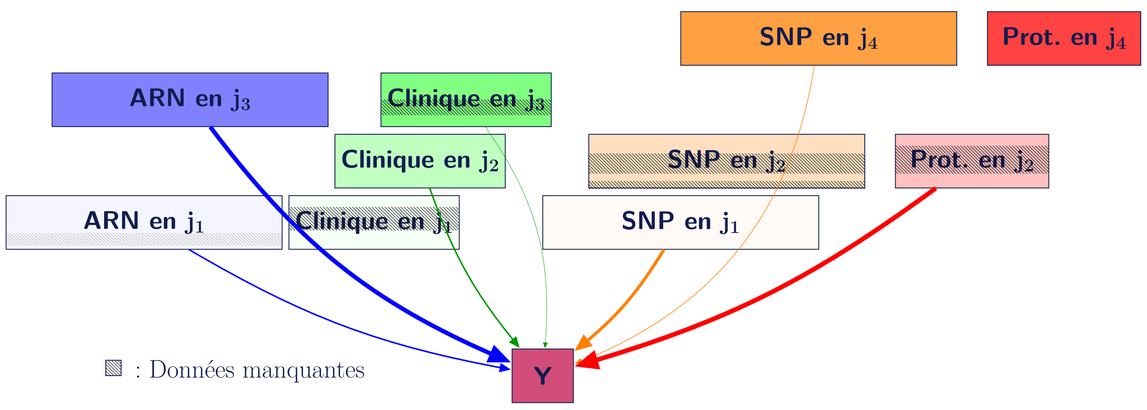

FIGURE – Structure multi-blocs supervisée avec échantillons manquants.

Le bloc à prédire, Y : détaille la question biologique. L'algorithme **Koh-Lanta** [4] gère les données manquantes d'*entraînement* et de *test*.

Simulations

Comparaison à des méthodes très utilisées :

- La moyenne, notée **mean**, très efficace malgré sa simplicité.
- **imputeMFA** [2], gère la structure par blocs.
- **softImpute** [1], très rapide et efficace si n grand.
- **nipals**, suggérée par **mixOmics** [3].

ddsPLS permet "imputation **&** prédiction". Comparaison à des approches en deux temps avec **Lasso** pour la prédiction,

a. Voir https://github.com/hlorenzo/ddsPLS, https://cran.r-project.org/package=ddsPLS et https://github.com/hlorenzo/py_ddspls.

FIGURE – Effet de la proportion de données manquantes sur l'erreur de prédiction RMSEP.

rVSV Ebola, Vaccin de phase 1

Un essai vaccinal de **phase 1**, 18 participants, 2 types de données hétérogènes.

Type	RNA-SEQ				Fonctionnalité Cellulaire			
Jour	0	1	3	7	0	1	3	7
$\#(variables)$	10279	10134	9082	9670	129	129	129	129

TABLE – Nombre de variables par bloc.

Le bloc à prédire, Y : 4 colonnes : valeurs d'anticorps plusieurs mois après vaccination. **ddsPLS** admet les meilleures performances.

Méthode		Jour 28	Jour 56	Jour 84	Jour 180	Mean
Imputation	Prédiction	RMSEP	RMSEP	RMSEP	RMSEP	RMSEP
ddsPLS		**1.027**	**0.6134**	**0.9426**	**1.029**	**0.9035**
Mean	**ddsPLS**	1.028	0.6312	1.041	**1.029**	0.9326
softImpute	**ddsPLS**	1.029±0	0.6326 ± 0.03795	1.027 ± 0.002191	**1.029±0.0001374**	0.9294
imputeMFA	**ddsPLS**	1.028	0.6899	1.026	**1.029**	0.9433

TABLE – Erreurs en validation croisée, Leave-One-OUt

L'interprétation, possible car méthode à sélection de variables :

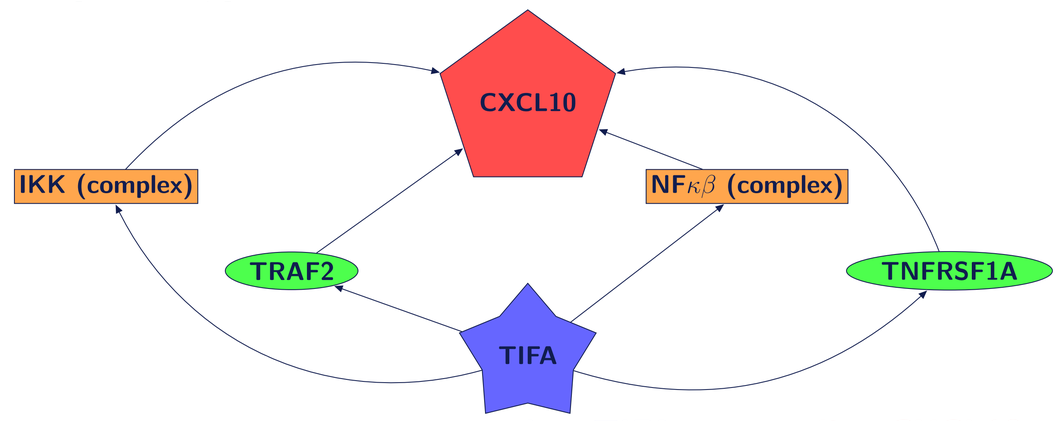

FIGURE – Pathways potentiels liant le **gène TIFA** et la **protéine CXCL10**

Le **gène TIFA** est associé à la **protéine CXCL10** via des gènes (TRAF2 et TNFRSF1A) et complexes protéiques (IKK et NF$\kappa\beta$). Data were analyzed through the use of IPA (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis).

Références

[1] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(1) :3367–3402, 2015.

[2] François Husson and Julie Josse. Handling missing values in multiple factor analysis. *Food quality and preference*, 30(2) :77–85, 2013.

[3] Kim-Anh Lê Cao, Ignacio González, and Sébastien Déjean. integromics : an r package to unravel relationships between two omics data sets. *Bioinformatics*, 25(21) :2855–2856, 2009.

[4] Hadrien Lorenzo, Jérôme Saracco, and Rodolphe Thiébaut. Supervised learning for multi-block incomplete data. *arXiv preprint arXiv :1901.04380*, 2019.

[5] Anne Rechtien, Laura Richert, Hadrien Lorenzo, ..., Rodolphe Thiébaut, Marcus Altfeld, and Marylyn Addo. Systems vaccinology identifies an early innate immune signature as a correlate of antibody responses to the ebola vaccine rvsv-zebov. *Cell Reports*, 20(9) :2251–2261, 09 2017.