Une PLS multi-blocs parcimonieuse pour données hétérogènes incomplètes

Hadrien Lorenzo[1], Jérôme Saracco[2], Rodolphe Thiébaut[1]

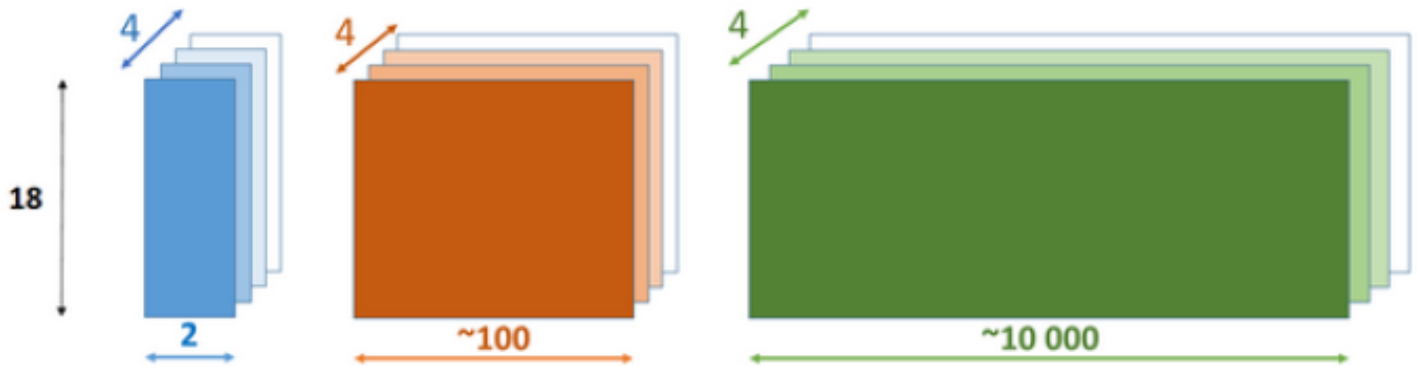
[1]SISTM (Inserm, U1219, Bordeaux Population Health and Inria, Talence, France) and Vaccine Research Institute, Creteil, France.
[2]CQFD (INRIA Bordeaux Sud-Ouest, France), CNRS (UMR5251)

JDS 2018, 19 novembre 2018

Motivation

Essai rVSV-ZEBOV Ebola de phase 1 avec doses échelonnées

- Premier vaccin à présenter une efficacité depuis la survenue de la maladie [Henao-Restrepo et al., *The Lancet*, 2017]



Réponse anticorps	Fonctionnalité cellulaire	Expression génétique
Jours 28, 56, 84, 180	Jours 0, 1, 3, 7	Jours 0, 1, 3, 7

Echantillons manquants : données génétiques

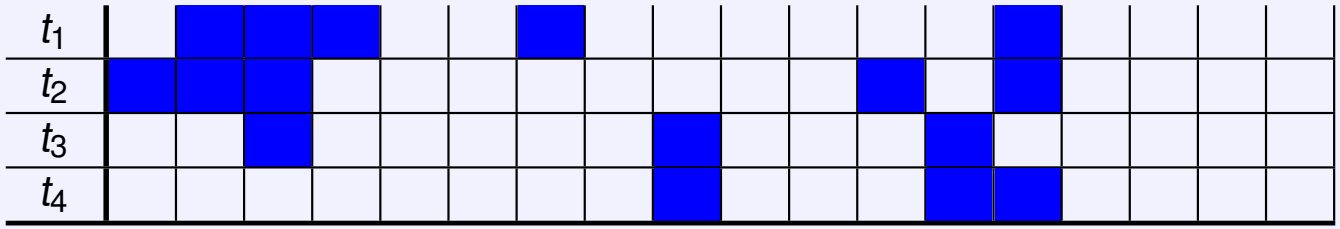


TABLE – Missing path du dataset Ebola rVSV-ZEBOV RNA-Seq où $t_1 = jour_0$, $t_2 = jour_1$, $t_3 = jour_3$ et $t_4 = jour_7$. Colonnes pour les participants.

- 30% de données/échantillons manquants,
- Lien "Missing structure"/"time structure"

Objectif

Prédire la réponse anticorps de façon parcimonieuse en gérant efficacement les données manquantes

Approches existantes

Modèle général

Combinent des alternances d'estimation :

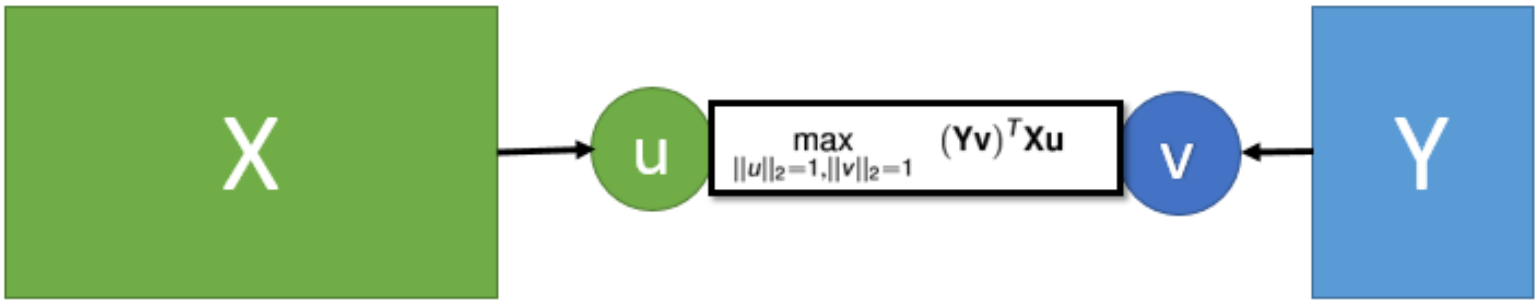
0. Initialiser les valeurs pour les données manquantes,
1. Estimer une factorisation des données complétées,
2. Estimer les données manquantes,
3. Recommencer en 1. jusqu'à convergence.

... en attente de stabilisation.
$\longrightarrow$ D'autant plus vrai dans le cas de modèles parcimonieux.
Côté utilisateur : difficile à optimiser

Contrainte majeure

Très peu d'individus : la stabilisation est plus difficile à trouver.

Approches PLS [Wold père et fils, 1983]



Équivalent à une recherche de sous-espaces propres (SVD). On appelle :

- **Poids** ou **weights** ou **loadings** u et v : importance donnée d'une variable de X, via u, et de Y, via v.
- **Scores** ou **variates** Xu et Yv : projections de X et de Y dans les sous-espaces définis par u et v.

$\Longrightarrow$ Rechercher dans X l'information qui est très liée à Y.

Résolution du problème de PLS

Utilisation du formalisme lagrangien :

$$\max_{u,v,\alpha_x,\alpha_y} (\mathbf{Y}v)^T \mathbf{X}u - \alpha_x/2(||u||_2^2 - 1) - \alpha_y/2(||v||_2^2 - 1),$$

$\mathbf{X}$ et $\mathbf{Y}$ les matrices échantillons, centrées, des covariables et des variables à prédire. α_x et α_y les coefficients de Lagrange. Alors :

Système :

$$\begin{cases} \partial_u . : & \alpha_x u = \mathbf{X}^T \mathbf{Y} v \\ \partial_v . : & \alpha_y v = \mathbf{Y}^T \mathbf{X} u \\ \partial_{\alpha_x} . : & ||u||_2^2 = 1 \\ \partial_{\alpha_y} . : & ||v||_2^2 = 1 \end{cases}$$

Optimisation :

1. $u \leftarrow \mathbf{X}^T \mathbf{Y} v$
2. $u \leftarrow u/||u||_2$
3. $v \leftarrow \mathbf{Y}^T \mathbf{X} u$
4. $v \leftarrow v/||v||_2$

Régression :

$$\begin{aligned} \mathbf{Y} &\approx \mathbf{XB} \\ \mathbf{B} &= \frac{v^T \mathbf{Y}^T \mathbf{X} u}{||\mathbf{X}u||_2^2} u v^T \end{aligned}$$

Classification :
LDA sur $(\mathbf{X}u, \mathbf{Y})$

Matrice de variance-covariance

Elle est au centre des approches PLS, via $\mathbf{Y}^T\mathbf{X}$!

La sélection de variables en PLS → sparse PLS

Principe, intérêt et pistes explorées

- Peu de mesures biologiques nécessaires en prédiction.
- Pénalisations $\mathcal{L}_1$ des poids

  $\Longrightarrow$ Sélection des variables et régularisation des données.

Des PLS parcimonieuses

- [Lê Cao et al., 2008], 2 paramètres/axe :

$$\min_{u,v} ||\mathbf{Y}^T\mathbf{X} - vu^T||_F^2 + \lambda_x||u||_1 + \lambda_y||v||_1$$

- [Chun et Keleş, 2010], $M = \mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}$, 3 paramètres/axe :

$$\min_{w,c} -\kappa w^T M w + (1-\kappa)(c-w)^T M (c-w) + \lambda_1||c||_1 + \lambda_2||c||_2$$
$$\text{subj. to } w^T w = 1,$$

Data Driven sPLS (**dd-sPLS**)

Idée

Travailler directement sur la matrice de variance-covariance en effaçant les liens entre les variables de X et de Y.

Solution utilisée :
Seuillage doux, S_λ, de $\frac{\mathbf{Y}^T\mathbf{X}}{n-1}$.

avec

$$S_\lambda(t) = \begin{cases} 0 & \text{si } |t| < \lambda \\ t - \lambda & \text{si } t \geqslant \lambda \\ t + \lambda & \text{si } t \leqslant -\lambda \end{cases}$$

Intêréts

- Sélectionner sur X et sur Y avec un seul paramètre, λ,
- Considérer un problème très bien connu : SVD

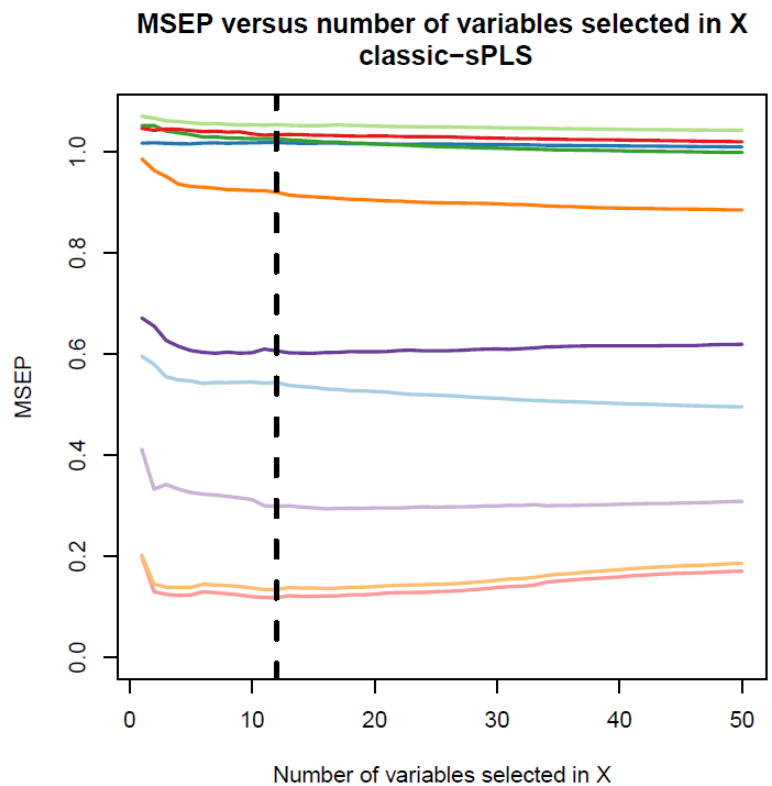
dd-sPLS

$$\max_{\substack{\mathbf{u}\in\mathbb{R}^{p\times R} \\ \mathbf{u}^T\mathbf{u}=\mathbb{I}_R}} ||S_\lambda\left(\frac{\mathbf{Y}^T\mathbf{X}}{n-1}\right)\mathbf{u}||_F^2,$$

Application : Liver Toxicity Dataset via classic sPLS

Voir [**heinloth2004gene**]. 64 Souris droguées. Expression de RNA. 10 variables réponses restituant l'état du foie, $\mathbf{X}_{64\times3116}$ and $\mathbf{Y}_{64\times10}$.
Comparaison **Classique sPLS** / **dd-sPLS**

- Paramètre de parcimonie en **Y** fixé arbitrairement à 2,
- Minimum d'erreur pour 12 covariables sélectionnées.



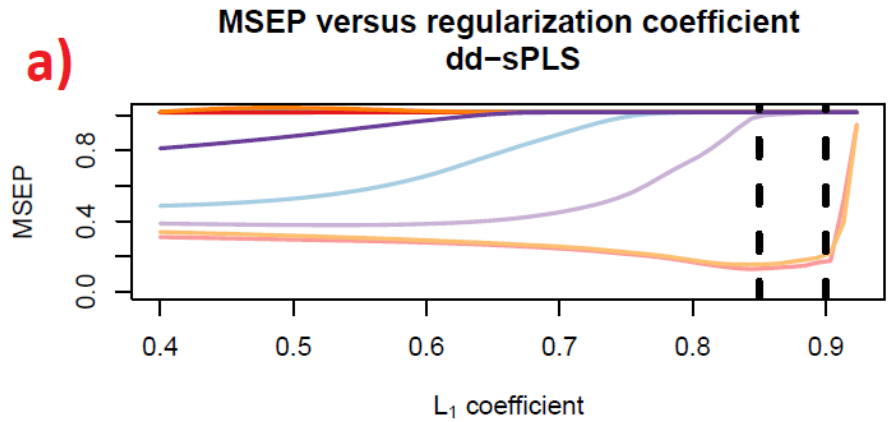
Application : Liver Toxicity Dataset via dd-sPLS

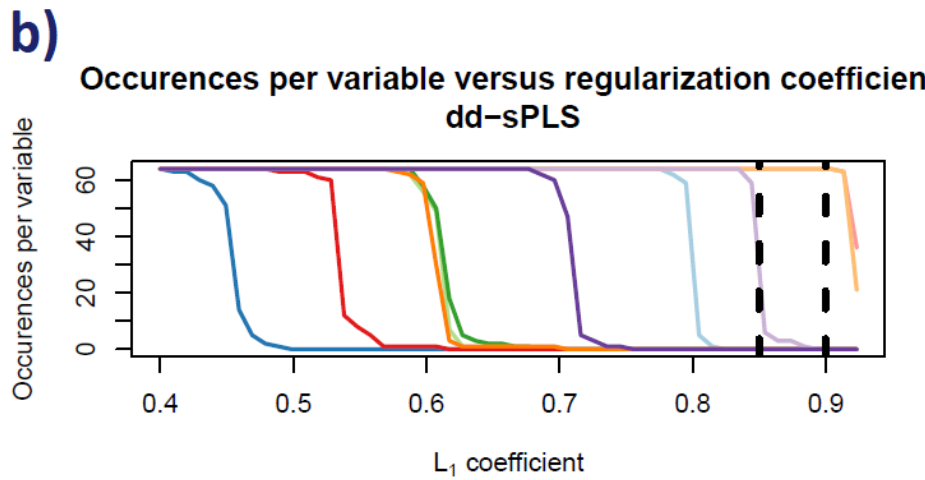
Deux visualisations disponibles :

- **a)** : L'erreur en prédiction,
- **b)** : Le nombre d'occurrences de chaque Y dans les modèles de validation

Observations :

- Via **a)** , $\lambda \approx 0.85$: 2 variables Y sélectionnées (?),
- Via **b)** : $\lambda \approx 0.9$ exactement 2 variables Y sélectionnées.





Liver Toxicity : Variables sélectionnées dans X

Classic-sPLS $keep_X = 12$		dd-sPLS $\lambda = 0.85$		dd-sPLS $\lambda = 0.9$	
Var	Coeff	Var	Coeff	Var	Coeff
A_43_P11724	0.172	A_43_P11724	0.172	**A_43_P14131**	**-0.862**
A_42_P802628	-0.117	A_42_P705413	-0.026	**A_42_P620915**	**-0.507**
A_43_P10606	-0.14	A_42_P802628	-0.117		
A_43_P14131	**-0.6**	A_43_P10606	-0.14		
A_42_P675890	-0.175	A_43_P22616	-0.012		
A_43_P23376	-0.213	**A_43_P14131**	**-0.6**		
A_42_P620915	**-0.515**	A_42_P675890	-0.175		
A_42_P758454	-0.175	A_43_P23376	-0.213		
A_42_P578246	-0.143	**A_42_P620915**	**-0.515**		
A_43_P17415	-0.331	A_42_P758454	-0.175		
A_42_P610788	-0.072	A_42_P578246	-0.143		
A_42_P840776	-0.264	A_43_P17415	-0.331		
		A_42_P610788	-0.072		
		A_42_P840776	-0.264		

TABLE – Comparaison des résultats de l'analyse du jeu de données de Liver Toxicity

Cas multi-blocs : mdd-sPLS avec données manquantes

Alternance de 2 étapes :
Construction du modèle et estimation des données manquantes.

Construction du modèle : Solution en deux étapes

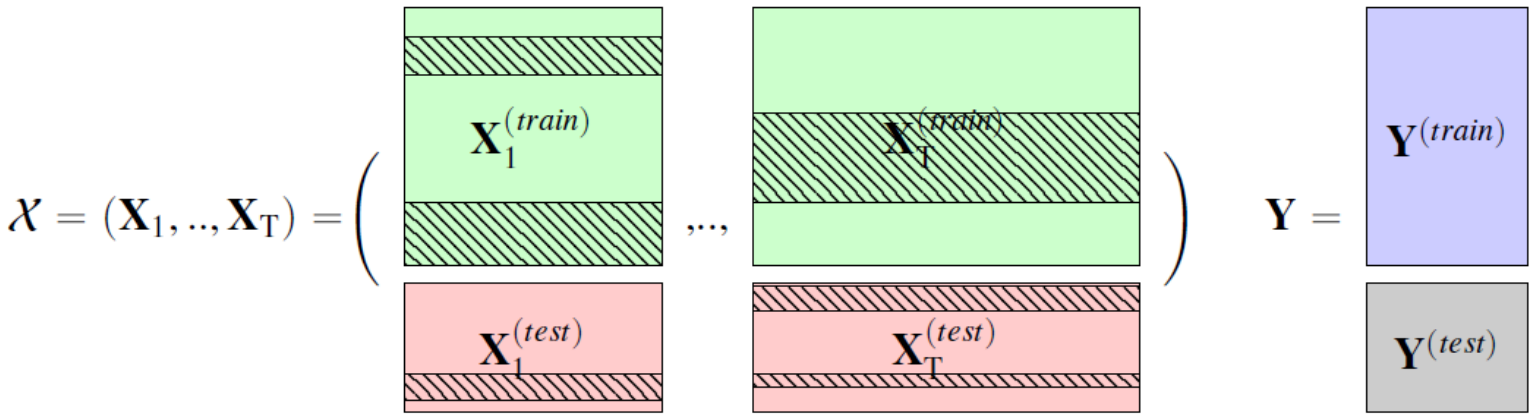
$$\max_{u_t^{(r)},\beta_t^{(r)}} \sum_{t=1}^{T}\sum_{r=1}^{R} {\beta_t^{(r)}}^2 ||S_\lambda\left(\frac{\mathbf{Y}^T\mathbf{X}_t}{n-1}\right)u_t^{(r)}||_2^2 \qquad (1)$$

1. $\forall t = 1..T,\ (u_t^{(1)},..,u_t^{(R)}) = \arg\max_{\mathbf{u}^T\mathbf{u}=\mathbb{I}_R} ||S_\lambda\left(\frac{\mathbf{Y}^T\mathbf{X}_t}{n-1}\right)\mathbf{u}||_F^2,$
2. Résolution de (1) via une SVD de R composantes

Estimation des données manquantes

Ceci grâce au modèle précédemment construit et à ce qui suit...

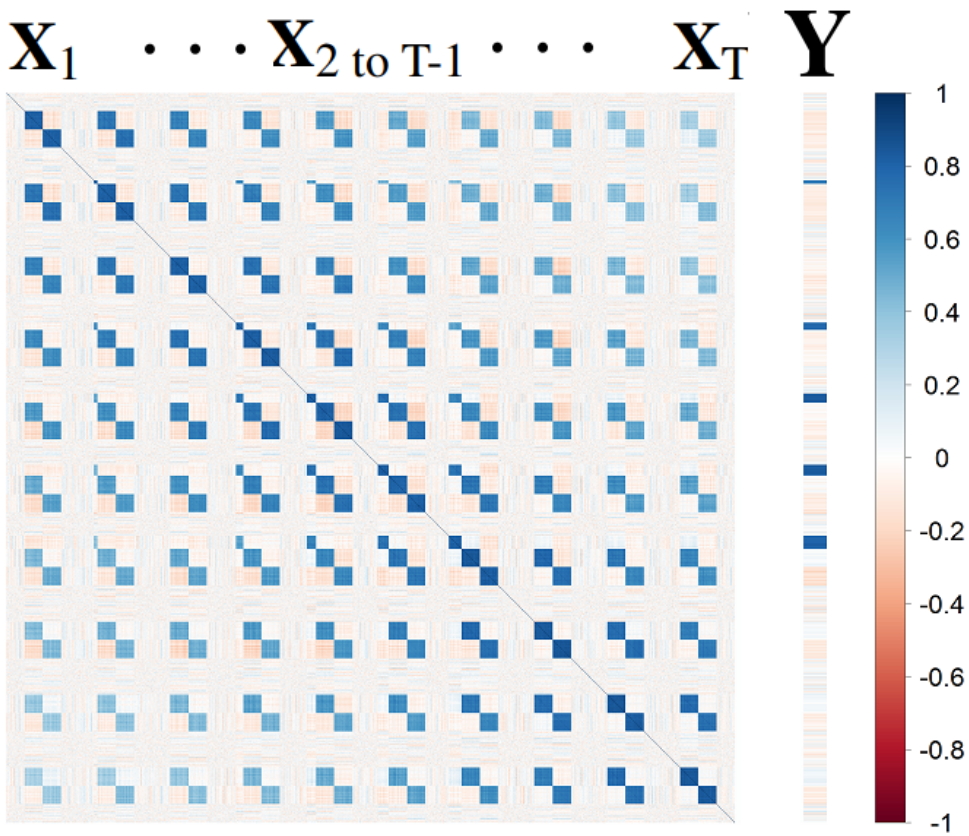
Gestion des données manquantes



- Le dataset d'**apprentissage** est imputé par bouclage sur lui-même (sans résultat théorique pour l'instant sur la convergence).
- Le dataset de **test** est imputé sans bouclage car l'apprentissage se fait sur la partie d'**apprentissage**.

Simulations

Construire un data-set de T blocs corrélé entre eux avec certaines variables de certains blocs corrélés à une variable réponse, univariée pour la comparaison à d'autres méthodes.



Résultats de simulations

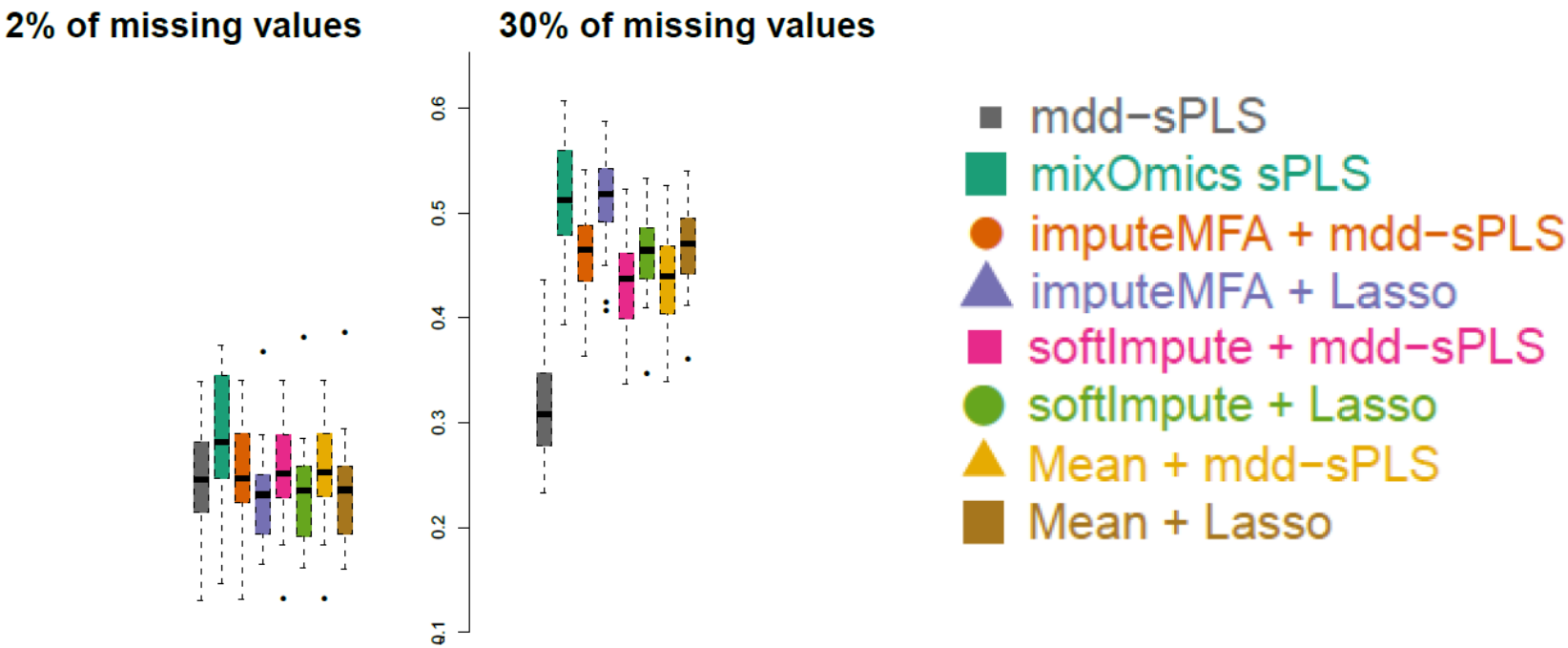
Comparaisons à des approches en 2 temps :

- Imputation :
  - missMDA (PCA + a-priori de groupe [Husson et Josse, 2013])
  - softImpute (modèle de prédiction [Hastie et al., 2015])
  - variable mean value
  - `nipals` fonction de mixOmics (+ sPLS classique).
- Prédiction sur données complétées : **mdd-sPLS** et **Lasso**

Protocole similaire à [Che et al., 2018] mais pour de la régression.
Challenge de réseaux récurrents.

Résultats de simulation

20 échantillons de 100 individus pour 10 blocs de 160 variables avec trois dimensions dont une seule est corrélée avec la réponse univariée.



Conclusion et futurs travaux

- Bons résultats de la méthode pour $p_{NA} \geqslant 20\%$,
- Etude du comportement, prédiction, sélection, convergence, pour n plus faible,
- Application aux données Ebola rVSV.

Merci !

References


Zhengping Che et al. “Recurrent neural networks for multivariate time series with missing values”. In : *Scientific reports* 8.1 (2018), p. 6085.


Hyonho Chun et Sündüz Keleş. “Sparse partial least squares regression for simultaneous dimension reduction and variable selection”. In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 72.1 (2010), p. 3–25.


Yash Deshpande et Andrea Montanari. “Sparse PCA via covariance thresholding”. In : *Advances in Neural Information Processing Systems*. 2014, p. 334–342.


Trevor Hastie et al. “Matrix completion and low-rank svd via fast alternating least squares”. In : *J. Mach. Learn. Res* 16.1 (2015), p. 3367–3402.


Ana Maria Henao-Restrepo et al. “Efficacy and effectiveness of an rVSV-vectored vaccine in preventing Ebola virus disease : final results from the Guinea ring vaccination, open-label, cluster-randomised trial (Ebola Ça Suffit !)” In : *The Lancet* 389.10068 (2017), p. 505–518.


François Husson et Julie Josse. “Handling missing values in multiple factor analysis”. In : *Food quality and preference* 30.2 (2013), p. 77–85.


Kim-Anh Lê Cao et al. “A sparse PLS for variable selection when integrating omics data”. In : *Statistical applications in genetics and molecular biology* 7.1 (2008).