# Une PLS multivoie parcimonieuse avec observations manquantes pour données hétérogènes

Hadrien Lorenzo [1] & Jérôme Saracco [2] & Rodolphe Thiébaut [3]

[1] *SISTM (Inserm, U1219, Bordeaux Population Health and Inria, Talence, France) and Vaccine Research Institute, Creteil, France*
hadrien.lorenzo@u-bordeaux.fr
[2] *CQFD - Inria Bordeaux Sud Ouest & IMB, ENSC - Bordeaux INP, 109 Avenue Roul, 33400 Talence, France*
jerome.saracco@ensc.fr
[3] *SISTM (Inserm, U1219, Bordeaux Population Health and Inria, Talence, France) and Vaccine Research Institute, Creteil, France*
rodolphe.thiebaut@u-bordeaux.fr

**Résumé.** Le problème de la régression sur données multivoie en grande dimension en présence d'observations manquantes est étudié. Cette méthode permet de sélectionner un nombre différent de variables pour chaque voie de façon cohérente, au travers de seulement 2 paramètres de pénalisation qui sont le nombre maximal de variables à sélectionner par voie et la corrélation minimale entre une variable sélectionnée et la variable réponse. Ceci est couplé à une approche itérative de prédiction des observations manquantes. Cette approche est appliquée à l'analyse de données hétérogènes répétées dans le temps issues d'un essai vaccinal de phase 1 du vaccin rVSV-ZEBOV contre Ebola.

**Mots-clés.** Données multivoie, sélection de variables, parcimonie, réduction de dimension, régression PLS, données manquantes, données hétérogènes, génomique, médecine

**Abstract.** The problem of regression on multiway data in an high dimensionnal context with missing observations is studied. This method allows to select a number of variables per way adapted to the nature of the multiway structure through 2 parameters which are the maximum number of variables selected and the minimum of correlation between the current variable and the output. This is realized with an iterative prediction of the missing observations. This method is applied to the analysis of a phase I vaccine trial for the rVSV-ZEBOV vaccine against Ebola implying heterogeneous data : gene expression and cellular functionnality measured repeateadly over time.

**Keywords.** Multiway data, variable selection, sparse, dimension reduction, PLS regression, missing data, heterogeneous data, genomic, medicine

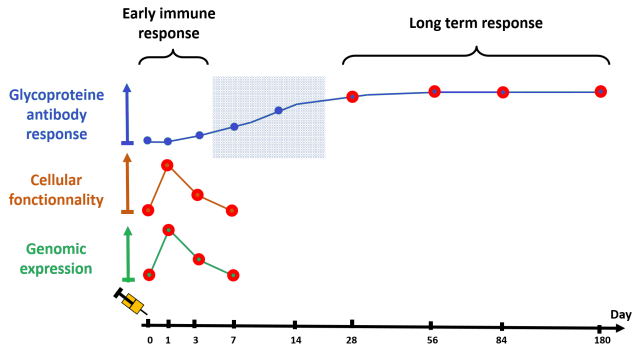
# 1 Introduction

Les données multivoie sont présentes dans de nombreux domaines et représentent $p$ variables quantitatives au travers de $T$ modalités, les voies, pour un même individu. Ces voies peuvent être de différentes natures. Dans le domaine de la santé, il peut s'agir de données d'imagerie, d'expression génique, de peptides... Il est aussi possible d'interpréter la longitudinalité des données comme une nouvelle dimension. Chaque temps est représenté par une voie. Ainsi en fonction de l'échantillonnage temporel, on a plus ou moins de voies. Nous supposerons acquises $T$ voies dans toute la suite de cette communication.
La problématique des données manquantes est ici abordée sans hypothése particulière sur le mécanisme à l'origine de ces non acquisitions. Un algorithme de type des moindres carrés alternés est mis en place de façon à faire ces estimations. De plus nous nous placerons dans le cas de problèmes supervisés par $q$ variables quantitatives.

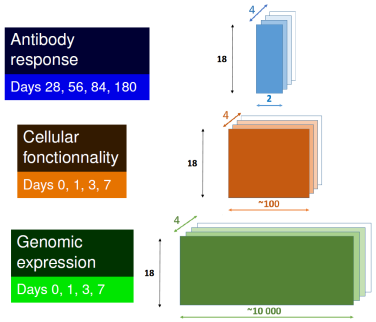
Dans la suite de cette communication nous commencerons par introduire les données et les problématiques qui en dcoulent avant de décrire notre méthode comme une généralisation de la PLS [Wold et al., 1983], pour *Partial Least Square*, comme la maximisation d'une covariance entre deux redescriptions de nos données dans l'espace des covariables et dans l'espace des variables à prédire.

# 2 Les données Ebola rVSV-ZEBOV de Hambourg

L'approche ici proposée a pour but de répondre aux problématiques posées par un jeu de données issu d'un essai vaccinal de phase I effectué à Hambourg et pour lequel les descriptions de $n = 18$ participants ont été recueillies. Les figures 1a et 1b récapitulent l'échantillonnage des données et leur structure générale.




(a) Allure générale des réponses des 18 participants à l'essai vaccinal




(b) Dimensions des différents jeux de données

Les données d'expression génique présentent des observations manquantes liées à une

mauvaise qualité de certains échantillons.

La question posée par ces données est de prédire le niveau d'anticorps à long terme en fonction de la réponse immunitaire précoce. Un premier travail [Rechtien et al., 2017] a permis de mettre en évidence l'efficacité de ce vaccin. Il a alors été possible de lier la réponse anticorps à la fonctionnalité cellulaire d'une part et à l'expression génétique d'autre part. La méthode PLS, voir par exemple [Wold et al., 1983], a été utilisée. Ceci a été réalisé en concaténant les tableaux de données, ce qui revient à ne pas prendre en compte leur structure longitudinale. On dit que le modèle est déplié.

L'approche proposée utilise une adaptation de la PLS à une structure multivoie de grande dimension avec observations manquantes. Une version parcimonieuse de cet outil a été utilisée dans un double but de régularisation et d'interprétabilité.

# 3 Méthode proposée

## 3.1 Cadre d'étude

Le modèle retenu est une généralisation de la méthode PLS appliquée à des variables latentes, elles-mêmes issues d'un problème de réduction de dimension. Il y a donc ici deux niveaux imbriqués de réduction de dimension, le premier qui réduit chaque voie à une variable latente et le second qui fait une régression sur ces variables latentes.

Soient la matrice $Y \in \mathcal{M}_{n\times q}(\mathbb{R})$ contenant les $q$ variables à prédire et $(X_t)_{t=1..T}$ les $T$ tableaux ($X_t \in \mathcal{M}_{n\times p}(\mathbb{R})$) contenant les $p$ prédicteurs et $X = [X_1, .., X_T]$ la concaténation des $X_t$. Dans la suite nous supposerons les matrices $X$ et $Y$ centrées, les variables de $X$ étant supposées de variance unitaire.

**Cas général.** La famille des problèmes d'optimisation considérée cherche à maximiser un critère de covariance entre une transformation de $X$ et une transformation de $Y$

$$\max_{(u_x,u_y)\in\mathbb{R}^{pT}\times\mathbb{R}^q} |f(X,u_x)^T g(Y,u_y)|, \tag{1}$$

où les fonctions considérées $f(.,u_x)$ et $g(.,u_y)$ sont à valeur dans $\mathbb{R}^n$. L'optimisation des paramètres $u_x$ et $u_y$ se fait souvent sous contraintes de normes, permettant une régularisation de la méthode. Le critère du problème (1) est, selon l'inégalité de Cauchy-Schwarz, majoré par $||f(X,u_x)||_2||g(Y,u_y)||_2$ avec cas d'égalité lorsque $g(Y,u_y)$ est colinéaire à $f(X,u_x)$. Ainsi s'il y a une bonne adéquation "données $\leftrightarrow$ modèle", on a :

$$\exists(\alpha,\epsilon) \in \mathbb{R}\times\mathbb{R}^n \text{ tel que } g(Y,u_y) = \alpha f(X,u_x) + \epsilon, \tag{2}$$

avec $\epsilon$ le terme résiduel. En supposant exclu le cas dégénéré $f(X,u_x) = 0$, une estimation de $\alpha$ est donnée par $\frac{f(X,u_x)^T g(Y,u_y)}{||f(X,u_x)||_2^2}$.

Dans un but de prédiction de $Y$, il convient maintenant d'approximer la fonction réciproque de $g(., u_y)$, que nous notons $g^\dagger(., u_x)$ :

$$Y \approx g^\dagger\Big(\frac{f(X,u_x)^T g(Y,u_y)}{||f(X,u_x)||_2^2} f(X,u_x) + \epsilon, u_y\Big). \tag{3}$$

**Cas particulier de la PLS.** Nous avons $f(X,u_x) = Xu_x$ et $g(Y,u_y) = Yu_y$ avec $||u_x||_2 = ||u_y||_2 = 1$. Il vient ainsi $\alpha_{PLS} = \frac{u_x^T X^T Y u_y}{||Xu_x||_2^2}$. Le problème ainsi posé est bien connu pour être celui de la décomposition SVD de $X^TY$. Ainsi (2) devient $Yu_y = \alpha X u_x + \epsilon$. En prenant $g^\dagger(., u_y) = .u_y^T$, il vient $g^\dagger(g(., u_y), u_y) = .u_y u_y^T$ qui est le projecteur sur le sous-espace engendré par $u_y$ et qui est donc une approximation pertinente de la fonction inverse de $g$. On peut alors réécrire (3) sous la forme $Y \approx Yu_y u_y^T = \alpha_{PLS} X u_x u_y^T + \epsilon u_y^T$. On retrouve alors la définition des coefficients $\beta_{PLS}$ de régression du modèle PLS, $Y = X\beta_{PLS} + \tilde{\epsilon}$ :

$$\beta_{PLS} = \alpha_{PLS} u_x u_y^T = \frac{u_x^T X^T Y u_y}{||Xu_x||_2^2} u_x u_y^T. \tag{4}$$

En pratique le terme résiduel $\tilde{\epsilon} := \epsilon u_y^T$ est diminué par résolutions successives de ce problème, retirant par déflation l'information déjà prise en compte.

**Cadre spécifique de notre modèle.** Nous considérons une composée de fonctions pour la description de $X$, soit $f(., u_x) = f_1(., u_{x,1}) \circ f_2(., u_{x,2})$ où $u_{x,1} = [u_{x,1}^{(1)}, .., u_{x,1}^{(T)}] \in \mathbb{R}^T$ et $u_{x,2} = diag([u_{x,2}^{(1)}, .., u_{x,2}^{(T)}]) \in \mathcal{D}_{Tp}(\mathbb{R})$ l'ensemble des matrices réelles diagonales d'ordre $Tp$. La fonction $f_2$ permet la réduction de dimension de chacune des voies $X_t$ en variables latentes, et la fonction $f_1$ combine les variables latentes obtenues. Comme en PLS nous avons choisi des fonctions linéaires :

$$f(X,u_x) = f_1(f_2(X,u_{x,2}), u_{x,1}) = \sum_{t=1}^{T} u_{x,1}^{(t)} X_t u_{x,2}^{(t)}, \tag{5}$$

Pour $Y$ nous avons opté pour un modèle linéaire $g(Y,u_y) = Yu_y$.

## 3.2 Contraintes choisies pour $u_{x,1}$, $u_{x,2}$ et $u_y$

**Contraintes sur $u_y$ et $u_{x,1}$.** Tout comme en régression PLS classique nous avons contraint $u_y$ et $u_{x,1}$ à être de norme unité.

**Contraintes sur $u_{x,2}$.** Les contraintes sont ici plus spécifiques. En effet, nous sommes dans un contexte de données hétérogènes et de grande dimension où la nature biologique de l'étude impose la sélection d'ensembles de variables comme marqueurs pronostiques du corrélat de protection du vaccin.

— *Problème de données hétérogènes.* Chaque bloc $X_t$ est divisé par sa plus grande valeur propre $\sigma_1(X_t)$, comme conseillé dans le cas de l'Analyse Factorielle Multiple, voir par exemple [Husson and Josse, 2013].
— *Problème de sélection de variables.* Nous avons opté pour un seuillage doux comme celui proposé par [Lê Cao et al., 2008] : un paramètre, noté $keep_X$, contrôlant le nombre de variables sélectionnées.
— *Prise en compte simultannée de la sélection de variables pour données hétérogènes.* Les deux problématiques précédentes sont en réalité indissociables. En effet, l'expression d'un grand nombre de gènes peut être associée à l'expression de seulement quelques variables de fonctionnalité cellulaire, il convient donc de pouvoir particulariser le nombre de variables à sélectionner dans chaque bloc. Nous avons opté pour une solution "pauvre" en paramètres : $keep_X$ qui fixe le nombre maximal de variables de chaque voie à inclure dans le modèle et $\rho \in [0, 1]$ le seuil minimal de corrélation (en valeur absolue) entre une variable d'une voie donnée et la variable latente $Yu_y$. Ces deux paramètres sont fixés par l'utilisateur. Notons que quand $\rho$ est proche de 0 la méthode garde $keep_x$ variables par voie alors que lorsque $\rho$ est proche de 1, seules les variables les plus corrélées sont retenues dans le modèle. Ceci induit donc implicitement une parcimonie par voie, observée en pratique. Afin de contrôler la variance des variables latentes ainsi construites, chacune d'elle a été divisée par le nombre de variables conservées.

## 3.3 Modèle de régression

Dans notre cadre spécifique d'étude, l'approximation de la fonction réciproque de $g(., u_y)$ choisie est $g^\dagger(., u_y) = .u_y^T$. Ainsi en reprenant (3) :

$$Y \approx \sum_{t=1}^{T} u_{x,1}^{(t)} X_t u_{x,2}^{(t)} u_y^T + \epsilon u_y^T. \tag{6}$$

On remarque qu'il s'agit d'un modèle de régression sur les variables latentes de chaque $X_t$. En notant alors $l = [l_1, .., l_t] = [X_1 u_{x,2}^{(1)}, .., X_T u_{x,2}^{(T)}] \in \mathcal{M}_{n\times T}(\mathbb{R})$ la matrice issue de la concaténation des variables latentes $l_t$ et $\beta = [u_{x,1}^{(1)} u_y^T, .., u_{x,1}^{(T)} u_y^T] \in \mathcal{M}_{T\times q}(\mathbb{R})$, nous avons

$$Y \approx l\beta + \epsilon u_y^T. \tag{7}$$

## 3.4 Prise en compte des observations manquantes

En utilisant avantageusement les structures de covariances construites, il nous a été possible d'estimer, au cours de la procédure décrite précédemment, les positions des observations manquantes dans les sous-espaces construits, ce qui est fait de manière itérative en stabilisant les sous-espaces avant de ré-estimer les positions.

# 4 Mise en oeuvre

Le modèle précédent a été implémenté dans **R**, des simulations ont été réalisées ainsi que l'application aux données décrites en section 2. Les résultats seront exposés lors de la communication.

# Références


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

[Lê Cao et al., 2008] Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., and Besse, P. (2008). A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1).

[Rechtien et al., 2017] Rechtien, A., Richert, L., Lorenzo, H., Martrus, G., Hejblum, B., Dahlke, C., Kasonta, R., Zinser, M., Stubbe, H., Matschl, U., Lohse, A., Krähling, V., Eickmann, M., Becker, S., Agnandji, S. T., Krishna, S., Kremsner, P. G., Brosnahan, J. S., Bejon, P., Njuguna, P., Addo, M. M., Siegrist, C.-A., Huttner, A., Kieny, M.-P., Moorthy, V., Fast, P., Savarese, B., Lapujade, O., Thiébaut, R., Altfeld, M., and Addo, M. (2017). 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.

[Wold et al., 1983] Wold, S., Martens, H., and Wold, H. (1983). The multivariate calibration problem in chemistry solved by the pls method. In *Matrix pencils*, pages 286–293. Springer.