

A mono/multi-block sparse PLS for heterogeneous data with missing samples

Hadrien Lorenzo¹, Jérôme Saracco², Rodolphe Thiébaut¹

¹SISTM (Inserm, U1219, Bordeaux Population Health and Inria, Talence, France)
and Vaccine Research Institute, Creteil, France.

²CQFD (INRIA Bordeaux Sud-Ouest, France), CNRS (UMR5251)

ISPED seminar of biostatistics, July 5, 2018



rVSV-ZEBOV Ebola Vaccine phase I dose escalation trial

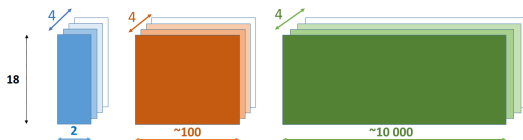
- First vaccine to show efficiency during the Ebola outbreak [Henao-Restrepo et al., *The Lancet*, 2017]

Hamburg vaccination dataset content

- 3 types of responses :
 - Antibody response
 - Cellular functionality
 - Genomic expression
- 18 participants divided in 2 vaccination groups :
 - $3 \cdot 10^6 pfu$
 - $20 \cdot 10^6 pfu$

rVSV-ZEBOV Ebola Vaccine phase I datasets

3 families of blocks of longitudinal data



Data analysis : high dimensional problem

$n = 18$, $p \in \{129, 18301\}$, 8 blocks ($T = 8$)

T : number of blocks \implies **multi-block** approach,
Variety of technologies \implies **heterogeneous data**.

Objective

Predict the **antibody response** (after months) with the immune response (after days). Unfolded analysis : forget temporal structure.

\rightarrow See [Rechtien et al., 2017]

Remaining big challenge : the missing values

Missing origins in the Genomic expression dataset

Poor sample qualities in case of :

- Low RNA integrity number (RIN)
- Insufficient library concentration
- Low sequencing depth

	7	5	9	1	15	10	14	4	2	12	17	16	8	18	13	11	3	6
t_1																		
t_2																		
t_3																		
t_4																		

Preliminar observations

- 30% of missing samples/values,
- Missing structure, parallel to time structure

⇒ Interest of a block structure

Existing solutions

Try many methods of imputations such as :

- 🔧 **Mean** imputation per variable per block,
- 🔧 **softImpute** [Hastie and Mazumder, 2015], no grouping structure
- 🔧 **missMDA** [Josse and Husson, 2016], variable grouping structure

Main problems of those methods :

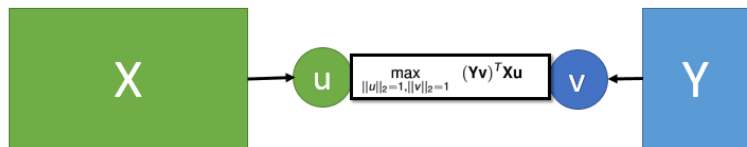
- 🔧 No variable selection,
- 🔧 Not converging,
- 🔧 Not supervised,
- 🔧 **Mean** is the best in that case.

Today : show you what we got!

A PLS-based method

- 🔧 Do variable selection,
- 🔧 Converges,
- 🔧 Is supervised,
- 🔧 Better than **Mean**

The PLS approaches, from [Wold father & son, 1983]



Equivalent to a eigen-space problem, or Singular Value Decomposition problem (**SVD**), with deflation. Under the common notations :

- 🏠 **Weights** or **loadings** or “**poids**” u and v : power given to a variable from X , via u , and from Y , via v .
- 🏠 **Scores** or **variates** of (**principal**) **components** Xu and Yv : projections of X and Y in the sub-spaces defined by u and v .

⇒ Research, by projections, in X the information linked to Y .

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

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

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. $v \leftarrow v / \|v\|_2$

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} &\leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{aligned}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

$$\left\{ \begin{array}{ll} \partial_{u \cdot} : & \alpha_x u = \mathbf{X}^T \mathbf{Y} v \\ \partial_{v \cdot} : & \alpha_y v = \mathbf{Y}^T \mathbf{X} u \\ \partial_{\alpha_x} : & \|u\|_2^2 = 1 \\ \partial_{\alpha_y} : & \|v\|_2^2 = 1 \end{array} \right. \quad \begin{array}{l} 1. \ u \leftarrow \mathbf{X}^T \mathbf{Y} v \\ 2. \ u \leftarrow u / \|u\|_2 \\ 3. \ v \leftarrow \mathbf{Y}^T \mathbf{X} u \\ 4. \ v \leftarrow v / \|v\|_2 \end{array}$$

$$\begin{array}{l} \mathbf{X} \leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{array}$$

Regression :

$$\mathbf{Y} \approx \mathbf{X} \mathbf{B} \\ \mathbf{B} = \frac{v^T \mathbf{Y}^T \mathbf{X} u}{\|\mathbf{X} u\|_2^2} u v^T$$

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

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

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

$$\begin{aligned} \mathbf{X} & \leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} & \leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{aligned}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

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

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. $v \leftarrow v / \|v\|_2$

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} &\leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{aligned}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_i = 0$:

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

Optimization (NIPALS) :

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. $v \leftarrow v / \|v\|_2$

Deflation :

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} &\leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{aligned}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

$$\left\{ \begin{array}{ll} \partial_{u \cdot} : & \alpha_x u = \mathbf{X}^T \mathbf{Y} v \\ \partial_{v \cdot} : & \alpha_y v = \mathbf{Y}^T \mathbf{X} u \\ \partial_{\alpha_x} : & \|u\|_2^2 = 1 \\ \partial_{\alpha_y} : & \|v\|_2^2 = 1 \end{array} \right. \quad \begin{array}{l} 1. \ u \leftarrow \mathbf{X}^T \mathbf{Y} v \\ 2. \ u \leftarrow u / \|u\|_2 \\ 3. \ v \leftarrow \mathbf{Y}^T \mathbf{X} u \\ 4. \ v \leftarrow v / \|v\|_2 \end{array}$$

$$\begin{array}{l} \mathbf{X} \leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{array}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

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

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. $v \leftarrow v / \|v\|_2$

$$\begin{aligned} \mathbf{X} &\leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} &\leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{aligned}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Resolution of the PLS problem

Under the \mathcal{L} agrangian formalism :

$$\max_{u,v,\alpha_x \geq 0, \alpha_y \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1),$$

$\mathbf{X}_{n \times p}$ and $\mathbf{Y}_{n \times q}$ the sample matrices, centered, of the covariates and of the response, then :

System $\partial_{\cdot} = 0$:

Optimization (NIPALS) :

Deflation :

$$\left\{ \begin{array}{ll} \partial_{u \cdot} : & \alpha_x u = \mathbf{X}^T \mathbf{Y} v \\ \partial_{v \cdot} : & \alpha_y v = \mathbf{Y}^T \mathbf{X} u \\ \partial_{\alpha_x} : & \|u\|_2^2 = 1 \\ \partial_{\alpha_y} : & \|v\|_2^2 = 1 \end{array} \right. \quad \begin{array}{l} 1. \ u \leftarrow \mathbf{X}^T \mathbf{Y} v \\ 2. \ u \leftarrow u / \|u\|_2 \\ 3. \ v \leftarrow \mathbf{Y}^T \mathbf{X} u \\ 4. \ v \leftarrow v / \|v\|_2 \end{array}$$

$$\begin{array}{l} \mathbf{X} \leftarrow \mathbf{X} - \mathbf{X} u u^T \\ \mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{Y} v v^T \end{array}$$

Regression :

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

Classification (PLS-DA) :

LDA on $(\mathbf{X} u, \mathbf{Y})$, u is built on the R successive components.

Variable selection in PLS → sparse PLS

Principle, interest and actual solutions

- Interest : Limit the number of biological measurements,
- Regularization shrinking \mathcal{L}_1 -norm of the weights, see [Tibshirani, 1996].

⇒ Selection & regularization.

Some sparse PLS

- [Lê Cao et al., 2008], 2 para./axis :

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

- [Chun and Keleş, 2010], $M = \mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$, 3 para./axis :

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

Variable selection in PLS → sparse PLS

Principle, interest and actual solutions

- Interest : Limit the number of biological measurements,
- Regularization shrinking \mathcal{L}_1 -norm of the weights, see [Tibshirani, 1996].

⇒ Selection & regularization.

Some sparse PLS

- [Lê Cao et al., 2008], **2 para./axis** :

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

- [Chun and Keleş, 2010], $M = \mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$, **3 para./axis** :

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

Variable selection in PLS → sparse PLS

Principle, interest and actual solutions

- Interest : Limit the number of biological measurements,
- Regularization shrinking \mathcal{L}_1 -norm of the weights, see [Tibshirani, 1996].

⇒ Selection & regularization.

Some sparse PLS

- [Lê Cao et al., 2008], **2 para./axis** :

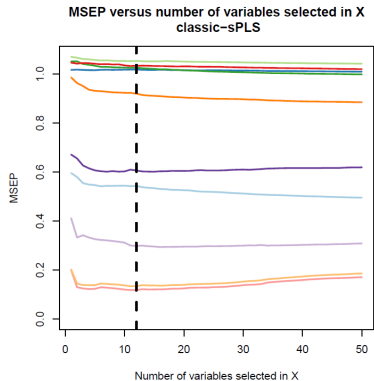
$$\min_{u,v} ||\mathbf{Y}^T \mathbf{X} - vu^T||_F^2 + \lambda_x ||u||_1 + \lambda_y ||v||_1$$

- [Chun and Keleş, 2010], $M = \mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$, **3 para./axis** :

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

Application : Liver Toxicity Dataset via classical sPLS [Lê Cao et al., 2008]

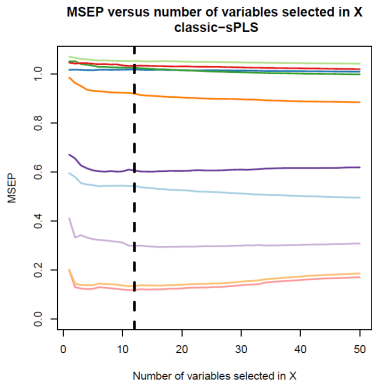
From [Heinloth et al., 2004]. 64 drugged mice and their RNA expression, 10 response variables about liver : $\mathbf{X}_{64 \times 3116}$, $\mathbf{Y}_{64 \times 10}$.



- $\lambda_y = f(keep_y)$, $keep_y = 2$ fixed,
- Min of error : 12 select. var. in X .
PB : How many Y var. in the model ?
 $2?...3?...5?...6?... | keep_y = 2 \}$
- **Good prediction** :
Many errors minimized,
- **Bad selection** : ≥ 5 variables predicted $| keep_y = 2 \}$.

Application : Liver Toxicity Dataset via classical sPLS [Lê Cao et al., 2008]

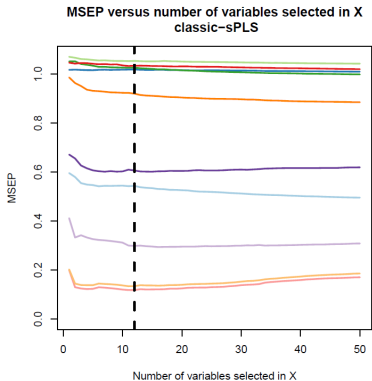
From [Heinloth et al., 2004]. 64 drugged mice and their RNA expression, 10 response variables about liver : $\mathbf{X}_{64 \times 3116}$, $\mathbf{Y}_{64 \times 10}$.



- $\lambda_y = f(keep_y)$, $keep_y = 2$ fixed,
- Min of error : 12 select. var. in X.
PB : How many Y var. in the model ?
 $2?...3?...5?...6?... | keep_y = 2\}$
- Good prediction :
Many errors minimized,
- Bad selection : ≥ 5 variables predicted $| keep_y = 2\}$.

Application : Liver Toxicity Dataset via classical sPLS [Lê Cao et al., 2008]

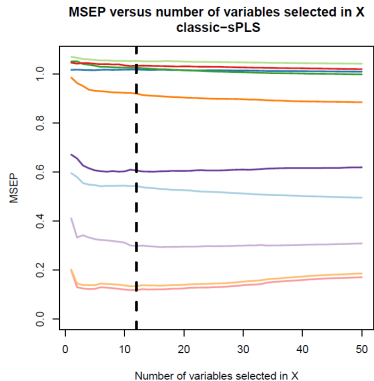
From [Heinloth et al., 2004]. 64 drugged mice and their RNA expression, 10 response variables about liver : $\mathbf{X}_{64 \times 3116}$, $\mathbf{Y}_{64 \times 10}$.



- $\lambda_y = f(keep_y)$, $keep_y = 2$ fixed,
- Min of error : 12 select. var. in X.
PB : How many Y var. in the model ?
 $2?...3?...5?...6?... | keep_y = 2\}$
- **Good prediction** :
Many errors minimized,
- **Bad selection** : ≥ 5 variables predicted $| keep_y = 2\}$.

Application : Liver Toxicity Dataset via classical sPLS [Lê Cao et al., 2008]

From [Heinloth et al., 2004]. 64 drugged mice and their RNA expression, 10 response variables about liver : $\mathbf{X}_{64 \times 3116}$, $\mathbf{Y}_{64 \times 10}$.



- $\lambda_y = f(keep_y)$, $keep_y = 2$ fixed,
- Min of error : 12 select. var. in X.
PB : How many Y var. in the model ?
 $2?...3?...5?...6?... | keep_y = 2\}$
- **Good prediction** :
Many errors minimized,
- **Bad selection** : ≥ 5 variables predicted $| keep_y = 2\}$.

sparse PLS : Resolution of the classical problem

Under the \mathcal{L} agrangian formalism, (β_x, β_y) fixed by the user :

$$\max_{u, v, (\alpha_x, \alpha_y, \beta_x, \beta_y) \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1) - \beta_x \|u\|_1 - \beta_y \|v\|_1, \quad (1)$$

System :

$$\begin{cases} \partial_{u \cdot} : \alpha_x u = \mathbf{X}^T \mathbf{Y} v - \beta_x \text{sign}(u) \\ \partial_{v \cdot} : \alpha_y v = \mathbf{Y}^T \mathbf{X} u - \beta_y \text{sign}(v) \\ \partial_{\alpha_x \cdot} : \|u\|_2^2 = 1 \\ \partial_{\alpha_y \cdot} : \|v\|_2^2 = 1 \end{cases}$$

Optimization :

1. $u \leftarrow \mathbf{S}_{\beta_x}(\mathbf{X}^T \mathbf{Y} v)$
2. $u \leftarrow u / \|u\|_2$
3. $v \leftarrow \mathbf{S}_{\beta_y}(\mathbf{Y}^T \mathbf{X} u)$
4. $v \leftarrow v / \|v\|_2$

where $t \rightarrow S_\lambda(t)$ is the soft-thresholding function.

Our idea

Flip S_λ , a non linear function, and $v \rightarrow \mathbf{X}^T \mathbf{Y} v$ and $u \rightarrow \mathbf{Y}^T \mathbf{X} u$, linear functions with a common λ .

sparse PLS : Resolution of the classical problem

Under the \mathcal{L} agrangian formalism, (β_x, β_y) fixed by the user :

$$\max_{u, v, (\alpha_x, \alpha_y, \beta_x, \beta_y) \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1) - \beta_x \|u\|_1 - \beta_y \|v\|_1, \quad (1)$$

System :

$$\begin{cases} \partial_{u \cdot} : \alpha_x u = \mathbf{X}^T \mathbf{Y} v - \beta_x \text{sign}(u) \\ \partial_{v \cdot} : \alpha_y v = \mathbf{Y}^T \mathbf{X} u - \beta_y \text{sign}(v) \\ \partial_{\alpha_x \cdot} : \|u\|_2^2 = 1 \\ \partial_{\alpha_y \cdot} : \|v\|_2^2 = 1 \end{cases}$$

Optimization :

1. $u \leftarrow \mathbf{S}_{\beta_x}(\mathbf{X}^T \mathbf{Y} v)$
2. $u \leftarrow u / \|u\|_2$
3. $v \leftarrow \mathbf{S}_{\beta_y}(\mathbf{Y}^T \mathbf{X} u)$
4. $v \leftarrow v / \|v\|_2$

where $t \rightarrow S_\lambda(t)$ is the soft-thresholding function.

Our idea

Flip S_λ , a non linear function, and $v \rightarrow \mathbf{X}^T \mathbf{Y} v$ and $u \rightarrow \mathbf{Y}^T \mathbf{X} u$, linear functions with a common λ .

sparse PLS : Resolution of the classical problem

Under the \mathcal{L} agrangian formalism, (β_x, β_y) fixed by the user :

$$\max_{u, v, (\alpha_x, \alpha_y, \beta_x, \beta_y) \geq 0} v^T \mathbf{Y}^T \mathbf{X} u - \alpha_x / 2 (\|u\|_2^2 - 1) - \alpha_y / 2 (\|v\|_2^2 - 1) - \beta_x \|u\|_1 - \beta_y \|v\|_1, \quad (1)$$

System :

$$\begin{cases} \partial_{u \cdot} : \alpha_x u = \mathbf{X}^T \mathbf{Y} v - \beta_x \text{sign}(u) \\ \partial_{v \cdot} : \alpha_y v = \mathbf{Y}^T \mathbf{X} u - \beta_y \text{sign}(v) \\ \partial_{\alpha_x \cdot} : \|u\|_2^2 = 1 \\ \partial_{\alpha_y \cdot} : \|v\|_2^2 = 1 \end{cases}$$

Optimization :

1. $u \leftarrow \mathbf{S}_{\beta_x}(\mathbf{X}^T \mathbf{Y} v)$
2. $u \leftarrow u / \|u\|_2$
3. $v \leftarrow \mathbf{S}_{\beta_y}(\mathbf{Y}^T \mathbf{X} u)$
4. $v \leftarrow v / \|v\|_2$

where $t \rightarrow S_\lambda(t)$ is the soft-thresholding function.

Our idea



Flip S_λ , a non linear function, and $v \rightarrow \mathbf{X}^T \mathbf{Y} v$ and $u \rightarrow \mathbf{Y}^T \mathbf{X} u$, linear functions with a common λ .

sparse PLS : Resolution of the data-driven problem

Optimization :

1. $u \leftarrow \mathbf{S}_\lambda(\mathbf{X}^T \mathbf{Y} / (n - 1))v$
2. $u \leftarrow u / \|u\|_2$
3. $v \leftarrow \mathbf{S}_\lambda(\mathbf{X}^T \mathbf{Y} / (n - 1))u$
4. $v \leftarrow v / \|v\|_2$

Interests

-  Select on X and on Y with 1 parameter : λ ,
-  Interpret λ : correlation threshold if X and Y standardized.

dd-sPLS : data driven sPLS on R components

$$\mathbf{u} = \arg \max_{\substack{\mathbf{u} \in \mathbb{R}^{p \times R} \\ \mathbf{u}^T \mathbf{u} = \mathbf{I}_R}} \|S_\lambda \left(\frac{\mathbf{Y}^T \mathbf{X}}{n - 1} \right) \mathbf{u}\|_F^2, \quad \mathbf{v} = \left(\frac{S_\lambda(\mathbf{N})^T u^{(r)}}{\|S_\lambda(\mathbf{N})^T u^{(r)}\|_2} \right)_{r=1..R} \quad (2)$$

Regression : PLS of $(t = \mathbf{X}\mathbf{u}, s = \mathbf{Y}\mathbf{v}) \implies \text{scores}(\mathbf{u}, \mathbf{v})$,
 $\mathbf{a} = \text{diag}(\mathbf{a}^{(r)})_{r=1..R} | \mathbf{a}^{(r)} = \langle s\mathbf{v}^{(r)}, t\mathbf{u}^{(r)} \rangle / \|t\mathbf{u}^{(r)}\|_2^2$ then

$$\mathbf{Y} \approx \mathbf{X}\mathbf{B}, \quad \mathbf{B} = \mathbf{u}\mathbf{u}\mathbf{a}\mathbf{v}^T \mathbf{v}^T$$

sparse PLS : Resolution of the data-driven problem

Optimization :

1. $u \leftarrow \mathbf{S}_\lambda(\mathbf{X}^T \mathbf{Y} / (n-1))v$
2. $u \leftarrow u / \|u\|_2$
3. $v \leftarrow \mathbf{S}_\lambda(\mathbf{X}^T \mathbf{Y} / (n-1))u$
4. $v \leftarrow v / \|v\|_2$

Interests

- 📊 Select on X and on Y with **1 parameter** : λ ,
- 📊 Interpret λ : correlation threshold if \mathbf{X} and \mathbf{Y} standardized.

dd-sPLS : data driven sPLS on R components

$$\mathbf{u} = \arg \max_{\substack{\mathbf{u} \in \mathbb{R}^{p \times R} \\ \mathbf{u}^T \mathbf{u} = \mathbf{I}_R}} \|S_\lambda \left(\frac{\mathbf{Y}^T \mathbf{X}}{n-1} \right) \mathbf{u}\|_F^2, \quad \mathbf{v} = \left(\frac{S_\lambda(\mathbf{N})^T u^{(r)}}{\|S_\lambda(\mathbf{N})^T u^{(r)}\|_2} \right)_{r=1..R} \quad (2)$$

Regression : PLS of $(t = \mathbf{X}\mathbf{u}, s = \mathbf{Y}\mathbf{v}) \implies \text{scores}(\mathbf{u}, \mathbf{v})$,
 $\mathbf{a} = \text{diag}(\mathbf{a}^{(r)})_{r=1..R} | \mathbf{a}^{(r)} = \langle s\mathbf{v}^{(r)}, t\mathbf{u}^{(r)} \rangle / \|t\mathbf{u}^{(r)}\|_2^2$ then

$$\mathbf{Y} \approx \mathbf{X}\mathbf{B}, \quad \mathbf{B} = \mathbf{u}\mathbf{u}\mathbf{a}\mathbf{v}^T \mathbf{v}^T$$

sparse PLS : Resolution of the data-driven problem

Optimization :

1. $u \leftarrow \mathbf{S}_\lambda(\mathbf{X}^T \mathbf{Y} / (n - 1))v$
2. $u \leftarrow u / \|u\|_2$
3. $v \leftarrow \mathbf{S}_\lambda(\mathbf{X}^T \mathbf{Y} / (n - 1))u$
4. $v \leftarrow v / \|v\|_2$

Interests

- 📊 Select on X and on Y with **1 parameter** : λ ,
- 📊 Interpret λ : correlation threshold if \mathbf{X} and \mathbf{Y} standardized.

dd-sPLS : data driven sPLS on R components

$$\mathbf{u} = \arg \max_{\substack{\mathbf{u} \in \mathbb{R}^{p \times R} \\ \mathbf{u}^T \mathbf{u} = \mathbb{I}_R}} \|S_\lambda \left(\frac{\mathbf{Y}^T \mathbf{X}}{n - 1} \right) \mathbf{u}\|_F^2, \quad \mathbf{v} = \left(\frac{S_\lambda(\mathbf{N})^T u^{(r)}}{\|S_\lambda(\mathbf{N})^T u^{(r)}\|_2} \right)_{r=1..R} \quad (2)$$

Regression : PLS of $(t = \mathbf{X}\mathbf{u}, s = \mathbf{Y}\mathbf{v}) \implies \text{scores}(\mathbf{u}, \mathbf{v})$,
 $\alpha = \text{diag}(\alpha^{(r)})_{r=1..R} | \alpha^{(r)} = \langle s^{(r)}, t^{(r)} \rangle / \|t^{(r)}\|_2^2$ then

$$\mathbf{Y} \approx \mathbf{X}\mathbf{B}, \quad \mathbf{B} = \mathbf{u}\alpha\mathbf{v}^T$$

dd-sPLS, a few theoretical results

Proposition 1, where $\mathbf{N} = \mathbf{Y}^T \mathbf{X} / (n - 1)$:

$\mathcal{L} : \lambda \rightarrow \max\{\|S_\lambda(\mathbf{N})u\|_2^2 \mid u^T u = 1\}$, is decreasing on $[0, 1]$ and continuous on $[0, 1] - \{\|\mathbf{N}\|_\infty\}$.

Interpretation : $\lambda \in [0, 1]$, permits to control the information in common to \mathbf{X} and \mathbf{Y} to put in the model \rightarrow Regularization

Proposition 2, symmetric in u and v :

$\forall \lambda \in [0, 1]$, denoting $C_i^{(\lambda)}$ the i^{th} -column of $S_\lambda(\mathbf{N})$, $u = (u_i)_{i=1..p}$ sol. of (2) and $v = S_\lambda(\mathbf{N})^T u / \|S_\lambda(\mathbf{N})^T u\|_2$ then:
 $\forall i = 1..p : \quad u_i = 0 \iff \langle C_i^{(\lambda)}, v \rangle = 0$.

Interpretation : The problem implies sparsity and admits **Upper bounds** on u and v cardinalities, decreasing with λ .

dd-sPLS, a question of monotonicity

Is the cardinality monotonically decreasing per component ?

No, a counter-example :

$$\frac{\mathbf{Y}^T \mathbf{X}}{n-1} =$$

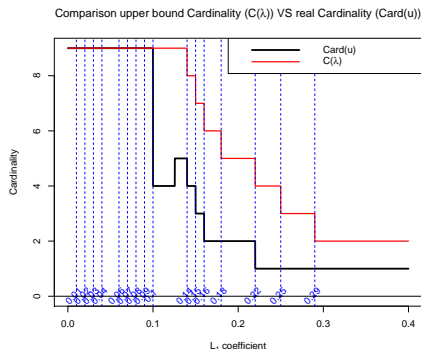
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Y_1	1.00	-0.06	-0.10	0.07	0.09	0.15	0.16	0.14	0.22
Y_2	-0.08	0.98	0.29	-0.18	0.25	0.02	0.04	-0.01	-0.03

2 components :

- Close in \mathcal{L}_2 -norm,
- Different in \mathcal{L}_0 -norm.

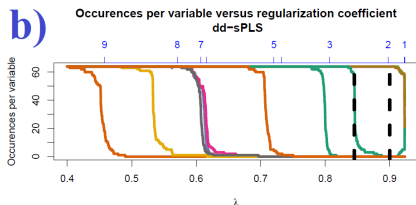
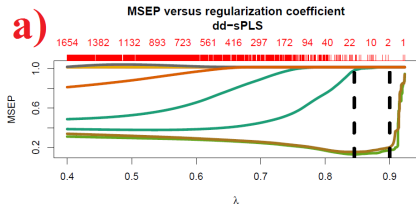
Reverse order in $\lambda \approx 0.13$.

Remark : Ordered through \mathcal{L}_2 -norm while \mathcal{L}_0 -norm is optimized in selection problems.



Application : Back to the Liver Toxicity Dataset

Results of the Cross-Validation



a) : MSEP,

b) : Selection per Y var.

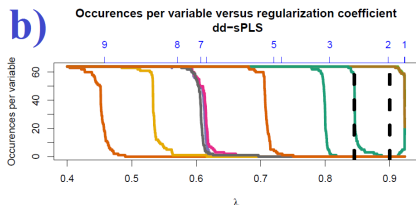
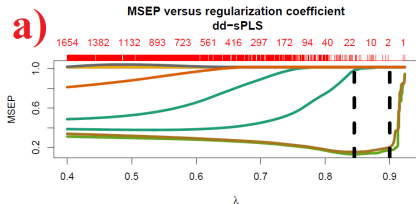
Observations

Via **a)**, $\lambda = 0.845$:
? 2 Y var. sel. ?

Via **b)** :
 $\lambda = 0.845$:
Exactly 2 Y var. sel.
 $\lambda \approx 0.9$:
Exactly 2 Y var. sel.

Application : Back to the Liver Toxicity Dataset

Results of the Cross-Validation



a) : MSEP,

b) : Selection per Y var.

Observations

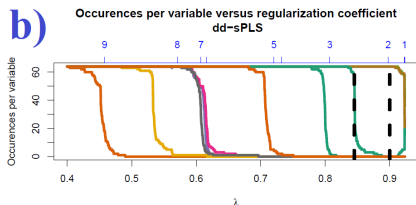
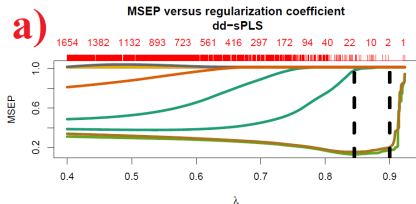
Via **a)** , $\lambda = 0.845$:
? 2 Y var. sel. ?

Via **b)** :

- ▶ $\lambda = 0.845$:
3rd Y var. sel. half times
- ▶ $\lambda \approx 0.9$:
Exactly 2 Y var. sel.

Application : Back to the Liver Toxicity Dataset

Results of the Cross-Validation



a) : MSEP,

b) : Selection per Y var.

Observations

Via **a)** , $\lambda = 0.845$:
? 2 Y var. sel. ?


Via **b)** :

- $\lambda = 0.845$:
3rd Y var. sel. half times
- $\lambda \approx 0.9$:
Exactly 2 Y var. sel.

Liver Toxicity Dataset : Comparison

Selection X variables comparison sPLS/dd-sPLS

Variable		A_43_P14131	A_42_P620915	A_43_P11724	A_42_P802628	A_43_P10606	A_42_P675890	A_43_P23376	A_42_P758454	A_42_P578246	A_43_P17415	A_42_P610788	A_42_P840776	A_42_P705413	A_43_P22616	Mean MSE(LOO)	Min MSE(LOO)
sPLS	$k_x = 12$	-0.6	-0.52	0.17	-0.12	-0.14	-0.18	-0.21	-0.18	-0.14	-0.33	-0.07	-0.26			0.65	0.11
dd-sPLS	$\lambda = 0.845$	-0.6	-0.52	0.17	-0.12	-0.14	-0.18	-0.21	-0.18	-0.14	-0.33	-0.07	-0.26	-0.03	-0.01	0.84	0.13
	$\lambda = 0.9$	-0.86	-0.51													0.85	0.17

 12 X var. sel. for classical sPLS. 15 in the case $\lambda = 0.845$ and 2 for $\lambda = 0.9$.

 Best **min** and **mean** errors for classical sPLS method.

Conclusion

dd-sPLS is better to select but worse to predict on that example.

Conclusion on the mono-block dd-sPLS

- Easy and well known problem (SVD),
- Selects X and Y variables with one parameter,
- Interpretable parameter : λ :

The minimum level of correlation between one X (or Y) variable and any of the Y (or X) variables to potentially get this variable in the model.

Multiblock PLS, called MBPLS

Formulation

Wold in 1984 [Wold, 1984] and Wangen & Kowalski [Wangen and Kowalski, 1989] consider T blocks indexed \mathbf{X}_t of predictors that can be bound to a response matrix \mathbf{Y} . Recalled weights u_t and scores $t_t = \mathbf{X}_t u_t$ for block \mathbf{X}_t , weight v and score $s = \mathbf{Y}v$ for \mathbf{Y} and finally super-weights $\mathbf{b} = (b_t)_{t=1..T}$ and

super-score $\mathbf{t} = \sum_{t=1}^T \mathbf{X}_t u_t b_t$ such as the 1^{st} component of the classical **MBPLS** maximizes :

$$\text{cov}^2(t, s) = \left(\sum_{t=1}^T v^T \mathbf{Y}^T \mathbf{X}_t u_t b_t \right)^2, \quad \text{subj. to } v^T v = u_t^T u_t = \mathbf{b}^T \mathbf{b} = 1 \quad (3)$$

Then deflation of \mathbf{X}_t 's and \mathbf{Y} and solves (3) anew, loop R times, R fixed by the user.

The deflation question

Component-wise method : solve sequential **MBPLS** with 2 cases of deflation in [Westerhuis and Smilde, 2001] :

- On each score : Poor prediction results,
- On the super-score : Better prediction results but mixing intra-block information.

→ Problem of variance restraining by outer axes. Thought shared with **François Husson** and **Arthur Tenenhaus**. **missMDA** [Josse and Husson, 2016] with no deflation and **RGCCA**, from [Tenenhaus and Tenenhaus, 2011], talk about a deflation-free solution.

⇒ No use of a deflation-based method.

mdd-sPLS : model definition

An (inter/intra)-blocks separable problem with no global iteration!

$$\arg \max_{(u_t^{(r)}, \beta_t^{(r)}) \in \mathbb{R}^{p_t} \times \mathbb{R}} \sum_{r=1}^R \sum_{t=1}^T \beta_t^{(r)2} \|S_\lambda \left(\frac{\mathbf{Y}^T \mathbf{X}_t}{n-1} \right) u_t^{(r)}\|_2^2 \quad \text{subj. to } \forall r, s | r \neq s \begin{cases} u_t^{(r)T} u_t^{(r)} = 1 \\ u_t^{(r)T} u_t^{(s)} = 0, \\ \sum_{t=1}^T \beta_t^{(r)2} = 1 \end{cases} \quad (4)$$

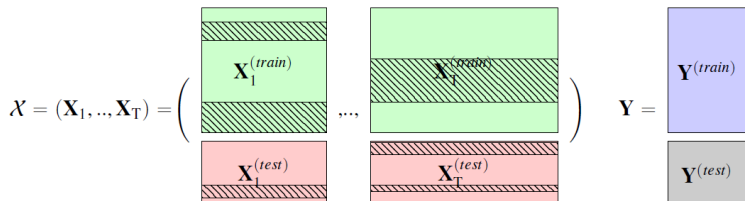
Inter-block : T independent dd-sPLS problems

$$\mathbf{u}_t = (u_t^{(1)}, \dots, u_t^{(R)}) = \arg \max_{\mathbf{u} \in \mathbb{R}^{p_t} \times r} \|\mathbf{M}_t(\lambda) \mathbf{u}\|_F^2, \quad \text{subj. to } \mathbf{u}^T \mathbf{u} = \mathbb{I}_R \quad (5)$$

Intra-block : R SVD problems

$$\beta^{(r)} = \arg \max_{\beta \in \mathbb{R}^T} \|z^{(r)}(\lambda) \beta\|_2^2, \quad \text{subj. to } \beta^T \beta = 1 \quad (6)$$

Missing data estimation : The *Koh-Lanta* algorithm



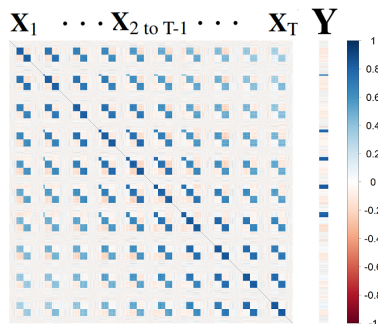
- The Tribe Stage : train** dataset imputation using mdd-sPLS prediction on s and λ . Using selected variables of global model : *Koh-Lanta* way of selection. Iterative process reestimating global model
- The Reunification Stage : test** dataset imputation, using mdd-sPLS prediction on t_{train} for non missing blocks and λ , on selected variables of main model. Non iterative process. Estimate Y_{test} reunifying all info.

Simulations

Build T -blocks data-set + Y matrix :

- Inter-block correlations : ρ_t ,
- Intra-block correlations : ρ_i ,
- Predictor/Response correlations : $\rho_{t\cdot}$

In each case define groups of variables with different sizes. Half of the blocks not linked to the response.



Chosen parameters

$T = 10$ blocks, 3 groups of variables, 40 variables per group & variable number of variables correlated to Y .

Baseline methods & question

2 step methods :

- Imputation : **missMDA**, **softImpute**, **Mean**, **nipals** (mixOmics solution),
- Prediction : **mdd-sPLS**, **Lasso** classical **sPLS** (for **nipals** imputation).

All-in-One method : [Che et al., *Scientific reports*, 2018], dealing with classification problems. Challenging recurrent neural networks. Huge n .

Simulation questions

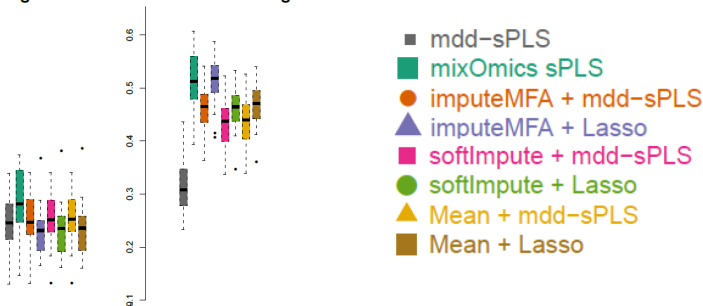
- Robustness to increasing number of missing values ?
- Robustness to low n and $n \ll p$?
- Robustness to low inter-block correlations ?

Robustness to increasing number of missing values ?

20 samples of **100 individuals** for **10 blocks** of **40 variables** each with **3 principal directions** where only **1** is correlated with the univariate response. $\rho_i = \rho_t = 0.9$. Mean Square Error (MSE).

2% of missing values

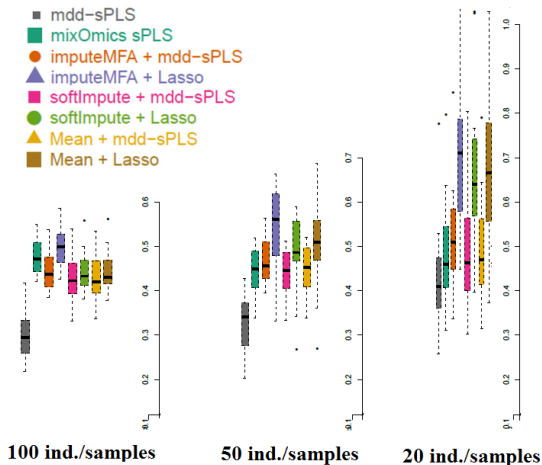
30% of missing values



The answer seems to be **Yes**.

Robustness to low n and $n \ll p$?

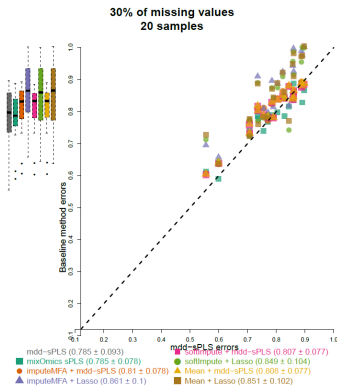
Change the number of individuals. MSE error



The answer seems to be **Yes**.

Robustness to Robustness to low inter-block correlations ?

$\rho_i = 0.9, \rho_t = 0.2$. MSE error



Hard for the all methods
Another type of simulations ?

Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$



mdd-sPLS

Gueckedou₂₈

Gueckedou₅₆

Gueckedou₈₄

Gueckedou₁₈₀

Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$



Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$

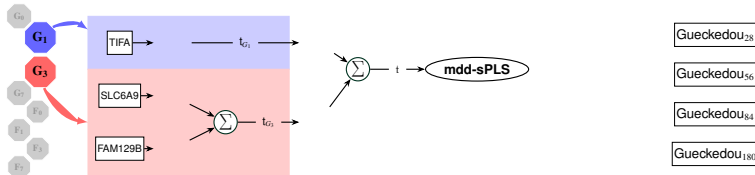


Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$

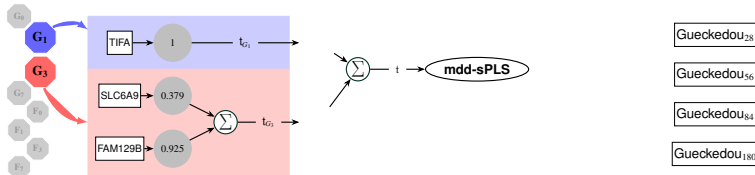


Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$

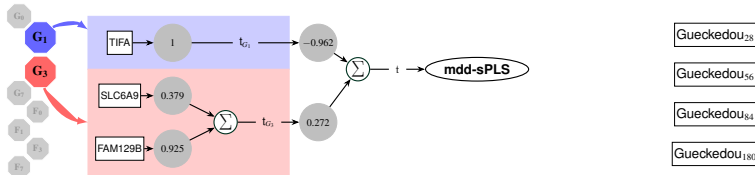


Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$

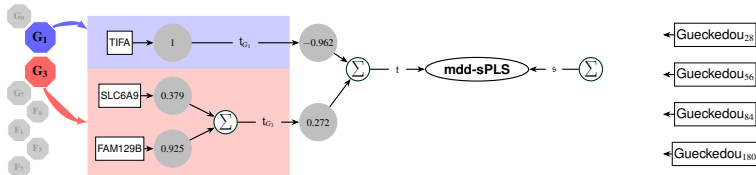


Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$

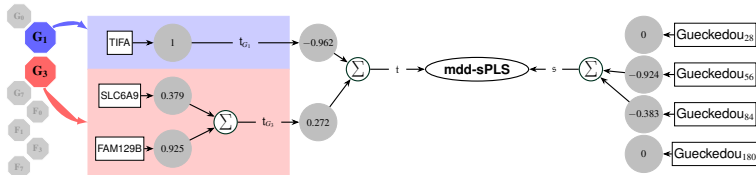


Application to the real data-set

Comparison Koh-Lanta/Mean imputation for dd-sPLS model

	Day 28		Day 56		Day 84		Day 180		Mean Error
	Error	#	Error	#	Error	#	Error	#	
Mean $\lambda \approx 0.863$	1.058	2	0.3985	18	1.084	6	1.059	0	0.8711
Koh-Lanta $\lambda \approx 0.865$	1.056	4	0.3796	18	0.9147	17	1.060	1	0.8318
Rel. gain (%)	0.19		4.7		16		-0.094		4.5

Final model : dd-sPLS with Koh-Lanta for $\lambda = 0.8653761$



Conclusion

dd-sPLS :

- Easy and well known problem (SVD),
- Selects X and Y variables with one parameter,
- Interpretable parameter : λ :

The minimum level of correlation between one X (or Y) variable and any of the Y (or X) variables to potentially get this variable in the model.

mdd-sPLS+Koh-Lanta :

- + dd-sPLS,
- Ok according to simulations,
- Works on real data,

Futur work :

- Test on new datasets,
- Publish + Finish package+vignette
- Create kernel dd-sPLS,

Thank you!

References I



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



Hyonho Chun and 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), pp. 3–25.



T Hastie and R Mazumder. “softImpute: Matrix Completion via Iterative Soft-Thresholded SVD”. In: *R package version 1* (2015).



Alexandra N Heinloth et al. “Gene expression profiling of rat livers reveals indicators of potential adverse effects”. In: *Toxicological Sciences* 80.1 (2004), pp. 193–202.



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), pp. 505–518.



Julie Josse and François Husson. “missMDA: a package for handling missing values in multivariate data analysis”. In: *Journal of Statistical Software* 70.1 (2016), pp. 1–31.

References II



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).



Anne Rechten et al. “Systems Vaccinology Identifies an Early Innate Immune Signature as a Correlate of Antibody Responses to the Ebola Vaccine rVSV-ZEBOV”. In: *Cell reports* 20.9 (2017), pp. 2251–2261.



Arthur Tenenhaus and Michel Tenenhaus. “Regularized generalized canonical correlation analysis”. In: *Psychometrika* 76.2 (2011), p. 257.



Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.



LE Wangen and BR Kowalski. “A multiblock partial least squares algorithm for investigating complex chemical systems”. In: *Journal of chemometrics* 3.1 (1989), pp. 3–20.



Johan A Westerhuis and Age K Smilde. “Deflation in multiblock PLS”. In: *Journal of chemometrics* 15.5 (2001), pp. 485–493.



S Wold. “Three PLS algorithms according to SW”. In: *Proc.: Symposium MULDAST (multivariate analysis in science and technology)*. 1984, pp. 26–30.

mdd-sPLS : Regression model

Objective and problem

$$\hat{\mathbf{Y}} = \sum_{t=1}^T \mathbf{X}_t \mathbf{B}_t,$$

Only $(\mathbf{X}_t \leftrightarrow \mathbf{Y})$ relations used :

\implies No adequacy between block components.

\implies Re-order components taking all info.

Solution : classical PLS solution on the super-scores

Denoting $\mathbf{b}_t = \text{diag}(\beta_t^{(1)}, \dots, \beta_t^{(R)})_{(R \times R)}$ the super-weights for each block, $\mathbf{t} = (\sum_{t=1}^T \mathbf{X}_t \mathbf{u}_t^{(r)} \beta_t^{(r)})_{r=1..R}$ and $\mathbf{s} = (\mathbf{Y} \mathbf{v}^{(r)})_{r=1..R}$:

$$\mathbf{B}_t = \mathbf{u}_t \mathbf{b}_t \mathbf{u}_t^T \mathbf{v}, \quad \left\{ \begin{array}{ll} (\mathbf{u}, \mathbf{v}) & : \text{Weights of PLS}(\mathbf{t}, \mathbf{s}) \\ \alpha & = \left(\frac{\langle \mathbf{s} \mathbf{v}^{(r)}, \mathbf{t} \mathbf{u}^{(r)} \rangle}{\|\mathbf{t} \mathbf{u}^{(r)}\|_2^2} \right)_{r=1..R} \end{array} \right.$$

Regularization path for rVSV-ZEBOV on mdd-sPLS

