Extension of Sparse PLS

B. Liquet

Incorporating Group structures within the data

- Natural example: Categorical variables which is a group of dummies variables in a regression setting.
- Genomics: genes within the same pathway have similar functions and act together in regulating a biological system.

$\hookrightarrow$ These genes can add up to have a larger effect

$\hookrightarrow$ can be detected as a group (i.e., at a pathway or gene set/module level).

We consider variables are divided into groups:

- {Example p: SNPs grouped into K genes} {

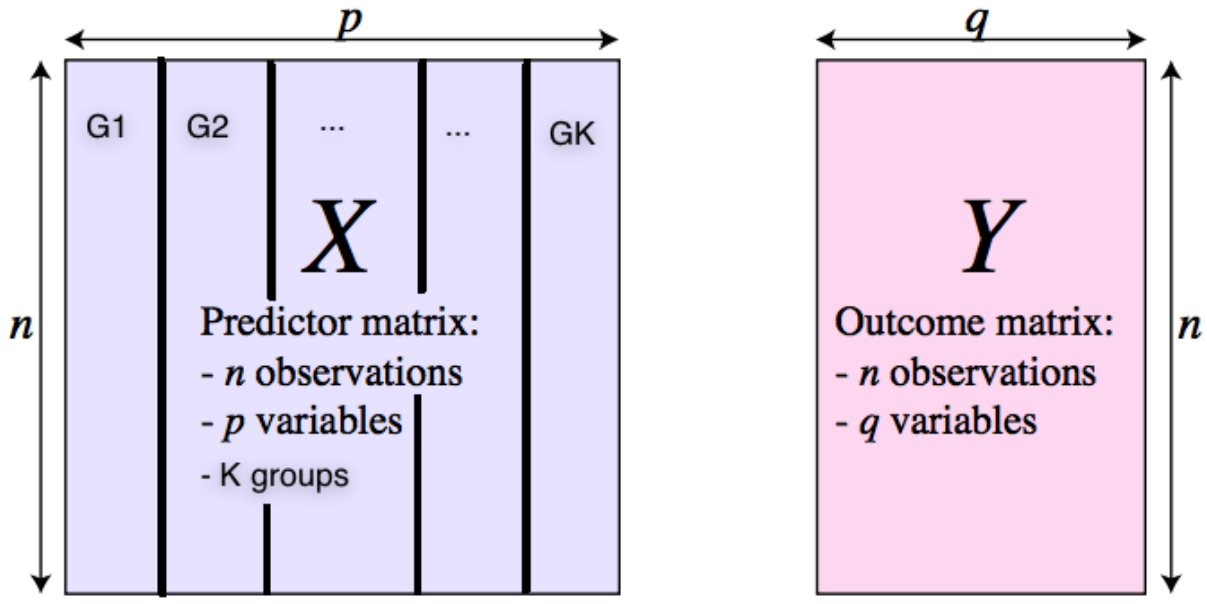
$$\mathbf{X} = [\underbrace{SNP_1, \ldots + SNP_k}_{gene_1} \mid \underbrace{SNP_{k+1}, SNP_{k+2}, \ldots, SNP_h}_{gene_2} \mid \ldots \mid \underbrace{SNP_{l+1}, \ldots, SNP_p}_{gene_K}]$$

}

- Example p: genes grouped into K pathways/modules ($X_j = \text{gene}_j$)

$$\mathbf{X} = [\underbrace{X_1, X_2, \ldots, X_k}_{M_1} \mid \underbrace{X_{k+1}, X_{k+2}, \ldots, X_h}_{M_2} \mid \ldots \mid \underbrace{X_{l+1}, X_{l+2}, \ldots, X_p}_{M_K}]$$

Aims in regression setting:



- Select group variables taking into account the data structures; all the variables within a group are selected otherwise none of them are selected
- Combine both sparsity of groups and within each group; only relevant variables within a group are selected

Sparse Models

Aim: Select gene expressions.

- sparse PLS

$$\xi = u_1 \times X_1 + 0 \times X_2 + u_3 \times X_3 + \cdots + u_p \times X_p$$

Aim: Select groups of gene expressions.

- group PLS

$$\xi = \underbrace{u_1 \times X_1 + u_2 \times X_2}_{Module\ 1} + \underbrace{0 \times X_3 + 0 \times X_4}_{Module\ 2} + \cdots + \underbrace{u_{p-1} \times X_{p-1} + u_p \times X_p}_{Module\ k}$$

Aim: Select group and within-group gene expressions.

- sparse group PLS

$$\xi = \underbrace{u_1 \times X_1 + 0 \times X_2}_{Module\ 1} + \underbrace{0 \times X_3 + 0 \times X_4}_{Module\ 2} + \cdots + \underbrace{u_{p-1} \times X_{p-1} + u_p \times X_p}_{Module\ k}$$

Optimisation functions: sPLS

Optimisation of the weights

- X-score $\xi = \mathbf{Xu}$, Y-score $\omega = \mathbf{Yv}$

$$\underset{\mathbf{v}_h^T\mathbf{v}_h \leqslant 1, \mathbf{u}_h^T\mathbf{u}_h \leqslant 1}{argmax} \quad Cov(\mathbf{Xu}, \mathbf{Yv}) - \lambda_1 ||\mathbf{u}||_1$$

- Sparse PLS

$$\xi = u_1 \times X_1 + 0 \times X_2 + u_3 \times X_3 + \cdots + u_p \times X_p$$

Sparse group PLS: gPLS

Optimisation of the weights

- X-score $\xi = \mathbf{Xu}$, Y-score $\omega = \mathbf{Yv}$

$$\underset{\mathbf{v}_h^\top \mathbf{v}_h \leqslant 1, \mathbf{u}_h^\top \mathbf{u}_h \leqslant 1}{argmax} \quad Cov(\mathbf{Xu}, \mathbf{Yv}) - \lambda_2 \sum_{k=1}^{K} \|\mathbf{u}^{(k)}\|_2$$

- Group PLS

$$\xi = \underbrace{0 \times X_1 + 0 \times X_2}_{Module\ 1} + \underbrace{0 \times X_3 + 0 \times X_4}_{Module\ 2} + \cdots + \underbrace{u_{p-1} \times X_{p-1} + u_p \times X_p}_{Module\ k}$$

Sparse Group PLS: sgPLS

Optimisation of the weights

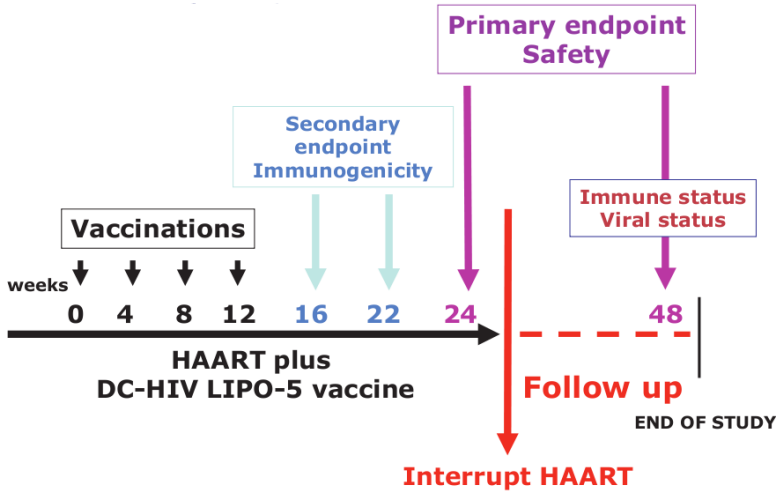
- X-score $\xi = \mathbf{Xu}$, Y-score $\omega = \mathbf{Yv}$

$$\underset{\mathbf{v}_h^\top \mathbf{v}_h \leqslant 1, \mathbf{u}_h^\top \mathbf{u}_h \leqslant 1}{argmax} \quad Cov(\mathbf{Xu}, \mathbf{Yv}) - \lambda_1 \|\mathbf{u}\|_1 - \lambda_2 \sum_{k=1}^{K} \|\mathbf{u}^{(k)}\|_2$$

- Sparse Group PLS

$$\xi = \underbrace{u_1 \times X_1 + 0 \times X_2}_{Module\ 1} + \underbrace{0 \times X_3 + 0 \times X_4}_{Module\ 2} + \cdots + \underbrace{u_{p-1} \times X_{p-1} + u_p \times X_p}_{Module\ k}$$

Illustration: DALIA trial



{

- Evaluation of the safety and the immunogenicity of a vaccine on $n = 19$ HIV-infected patients.
- The vaccine was injected on weeks 0, 4, 8 and 12 while patients received an antiretroviral therapy.
- An interruption of the antiretrovirals was performed at week 24.
- After vaccination, a deep evaluation of the immune response was performed at week 16.
- Repeated measurements of the main immune markers and gene expression were performed every 4 weeks until the end of the trials.

DALIA trial: Question ?

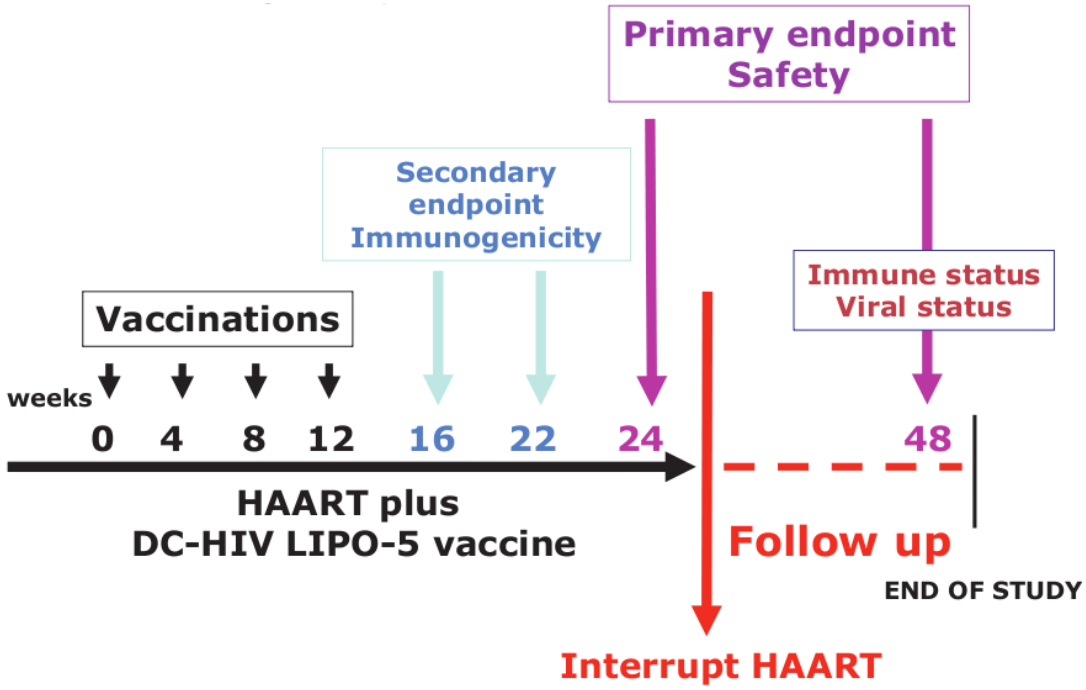
First results obtained using group of genes

- Significant change of gene expression among 69 modules over time before antiretroviral treatment interruption.

DALIA trial: Question ?

First results obtained using group of genes

- Significant change of gene expression among 69 modules over time before antiretroviral treatment interruption.
- How the gene abundance of these 69 modules as measured at week 16 **correlated** with immune markers measured at the same time.



sPLS, gPLS and sgPLS

- Responses variables **Y**= immune markers composed of $q = 7$ cytokines (IL21, IL2, IL13, IFNg, Luminex score, TH1 score, CD4).
- Predictors variables **X**= gene expressions ($p = 5399$) extracted from the 69 modules.
- Use the structure of the data (modules) for gPLS and sgPLS. Each gene belongs to one of the 69 modules.
- Asymmetric situation.

Results

- Tuning parameters: number of components, number of selected groups, number of selected genes
  ↪ mean square error of prediction (MSEP)
  ↪ estimated by K-fold cross-validation
- Cumulative percentage of variance of the responses:

  Table 1: Cumulative percentage of variance of the responses explained by the components for the sPLS, gPLS and sgPLS methods.

	comp1	comp2	comp3
sPLS	70.05	84.19	89.53
gPLS	55.13	73.72	83.43
sgPLS	64.18	83.19	89.25

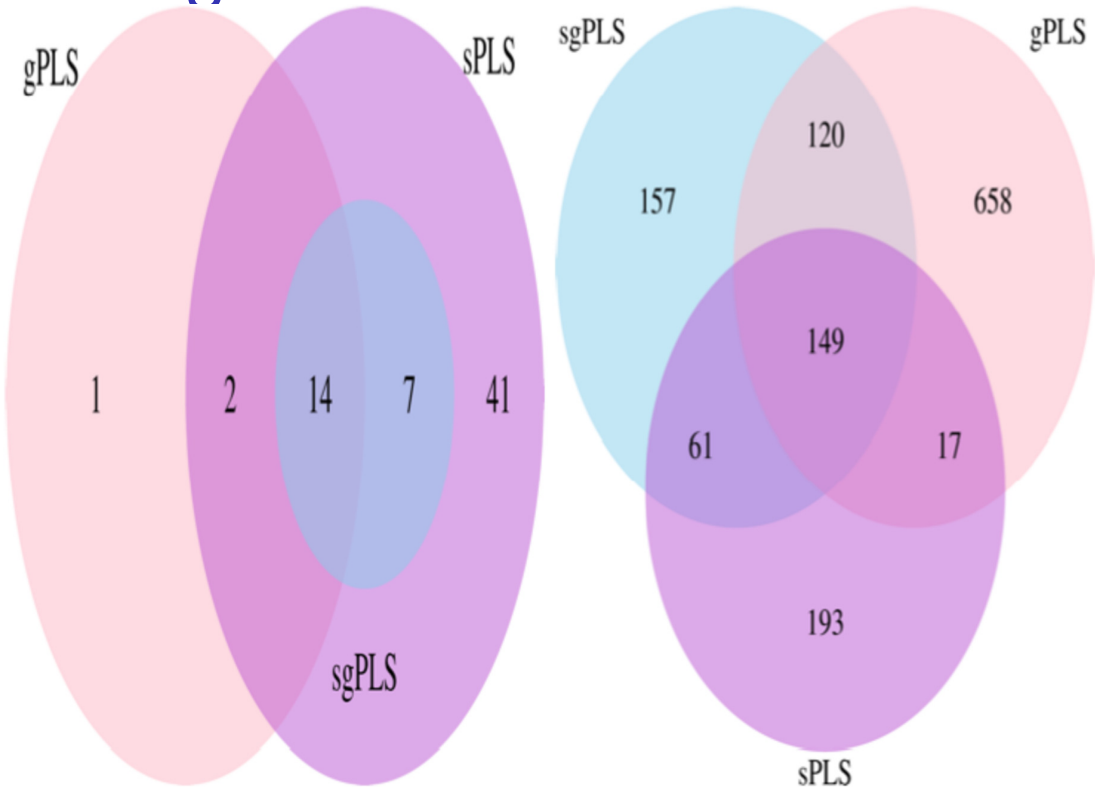
Results: Modules and number of genes selected

		gPLS			sgPLS			sPLS		
	size	comp1	comp2	comp3	comp1	comp2	comp3	comp1	comp2	comp3
M1.1	79	79	0	0	19	0	0	8	2	1
M3.2	126	126	0	0	41	0	0	22	0	0
M3.5	131	0	0	0	11	24	0	7	7	1
M3.6	42	42	0	0	15	0	0	6	0	0
M4.1	60	0	0	0	6	0	0	4	0	0
M4.13	72	72	0	0	26	0	0	11	0	0
M4.15	41	41	0	0	15	0	0	10	0	1
M4.2	43	43	0	0	14	0	0	7	1	1
M4.6	104	104	0	0	28	0	0	16	2	0
M5.1	214	0	0	0	46	0	0	21	2	4
M5.14	54	54	0	0	13	0	0	7	0	2
M5.15	24	24	24	0	20	0	0	18	0	0
M5.7	119	0	0	0	18	0	40	8	0	2
M6.13	38	38	0	0	10	0	0	7	0	0
M6.6	40	40	0	0	19	0	0	11	0	0
M7.1	150	150	0	0	37	0	0	19	2	2
M7.27	29	29	0	0	8	0	0	3	0	1
M4.7	82	0	0	0	0	20	0	5	7	0
M6.7	62	0	0	0	0	23	0	3	4	1
M8.59	13	0	13	0	0	4	0	0	3	0
M5.2	65	0	0	0	0	0	32	0	1	0
M4.8	53	53	0	0	0	0	0	1	0	0
M7.35	19	19	0	0	0	0	0	1	1	0
M4.11	17	0	0	17	0	0	0	0	0	0

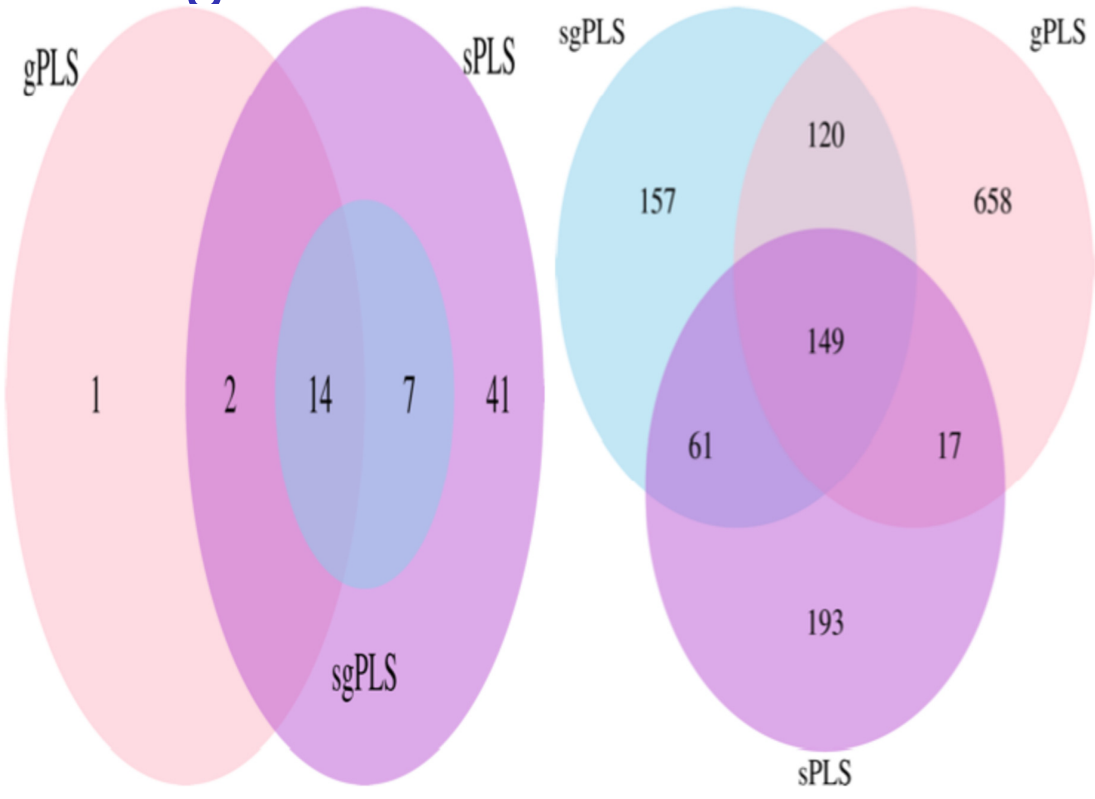
Results: Modules and number of genes selected

		gPLS			sgPLS			sPLS		
	size	comp1	comp2	comp3	comp1	comp2	comp3	comp1	comp2	comp3
M1.1	79	79	0	0	19	0	0	8	2	1
M3.2	126	126	0	0	41	0	0	22	0	0
M3.5	131	0	0	0	11	24	0	7	7	1
M3.6	42	42	0	0	15	0	0	6	0	0
M4.1	60	0	0	0	6	0	0	4	0	0
M4.13	72	72	0	0	26	0	0	11	0	0
M4.15	41	41	0	0	15	0	0	10	0	1
M4.2	43	43	0	0	14	0	0	7	1	1
M4.6	104	104	0	0	28	0	0	16	2	0
M5.1	214	0	0	0	46	0	0	21	2	4
M5.14	54	54	0	0	13	0	0	7	0	2
M5.15	24	24	24	0	20	0	0	18	0	0
M5.7	119	0	0	0	18	0	40	8	0	2
M6.13	38	38	0	0	10	0	0	7	0	0
M6.6	40	40	0	0	19	0	0	11	0	0
M7.1	150	150	0	0	37	0	0	19	2	2
M7.27	29	29	0	0	8	0	0	3	0	1
M4.7	82	0	0	0	0	20	0	5	7	0
M6.7	62	0	0	0	0	23	0	3	4	1
M8.59	13	0	13	0	0	4	0	0	3	0
M5.2	65	0	0	0	0	0	32	0	1	0
M4.8	53	53	0	0	0	0	0	1	0	0
M7.35	19	19	0	0	0	0	0	1	1	0
M4.11	17	0	0	17	0	0	0	0	0	0
M2.1	105	0	0	0	0	0	0	1	0	0
M3.1	74	0	0	0	0	0	0	1	0	0
M4.12	87	0	0	0	0	0	0	1	0	1
M4.16	79	0	0	0	0	0	0	2	0	1
M4.9	87	0	0	0	0	0	0	4	1	1
M5.10	196	0	0	0	0	0	0	3	3	0
M5.11	59	0	0	0	0	0	0	3	2	0
M5.13	147	0	0	0	0	0	0	1	2	4
M5.3	91	0	0	0	0	0	0	3	1	0
M5.4	115	0	0	0	0	0	0	3	2	2
M5.5	211	0	0	0	0	0	0	12	4	0
M5.6	126	0	0	0	0	0	0	3	2	1
M5.8	97	0	0	0	0	0	0	4	1	0
M5.9	72	0	0	0	0	0	0	4	0	0
M6.10	67	0	0	0	0	0	0	4	0	0
M6.14	33	0	0	0	0	0	0	3	0	0
M6.2	121	0	0	0	0	0	0	2	2	1
M6.20	42	0	0	0	0	0	0	1	2	0
M6.4	82	0	0	0	0	0	0	3	2	0
M6.9	35	0	0	0	0	0	0	2	1	0
M7.11	104	0	0	0	0	0	0	2	2	1
M7.12	108	0	0	0	0	0	0	4	0	0
M7.14	48	0	0	0	0	0	0	4	1	0
M7.15	78	0	0	0	0	0	0	2	0	1
M7.16	56	0	0	0	0	0	0	1	2	1
M7.2	93	0	0	0	0	0	0	4	1	0
M7.21	76	0	0	0	0	0	0	3	0	0
M7.24	65	0	0	0	0	0	0	2	0	0
M7.25	93	0	0	0	0	0	0	3	2	3
M7.26	63	0	0	0	0	0	0	2	0	0
M7.4	109	0	0	0	0	0	0	4	2	0
M7.5	132	0	0	0	0	0	0	6	5	2
M7.6	94	0	0	0	0	0	0	2	3	1
M7.8	85	0	0	0	0	0	0	3	0	0
M8.13	27	0	0	0	0	0	0	1	0	0
M8.14	27	0	0	0	0	0	0	2	1	0
M7.33	49	0	0	0	0	0	0	0	1	0
M7.7	89	0	0	0	0	0	0	0	3	1
M4.14	55	0	0	0	0	0	0	0	0	1
M4.4	68	0	0	0	0	0	0	0	0	1
M4.5	74	0	0	0	0	0	0	0	0	1

Results: Venn diagram



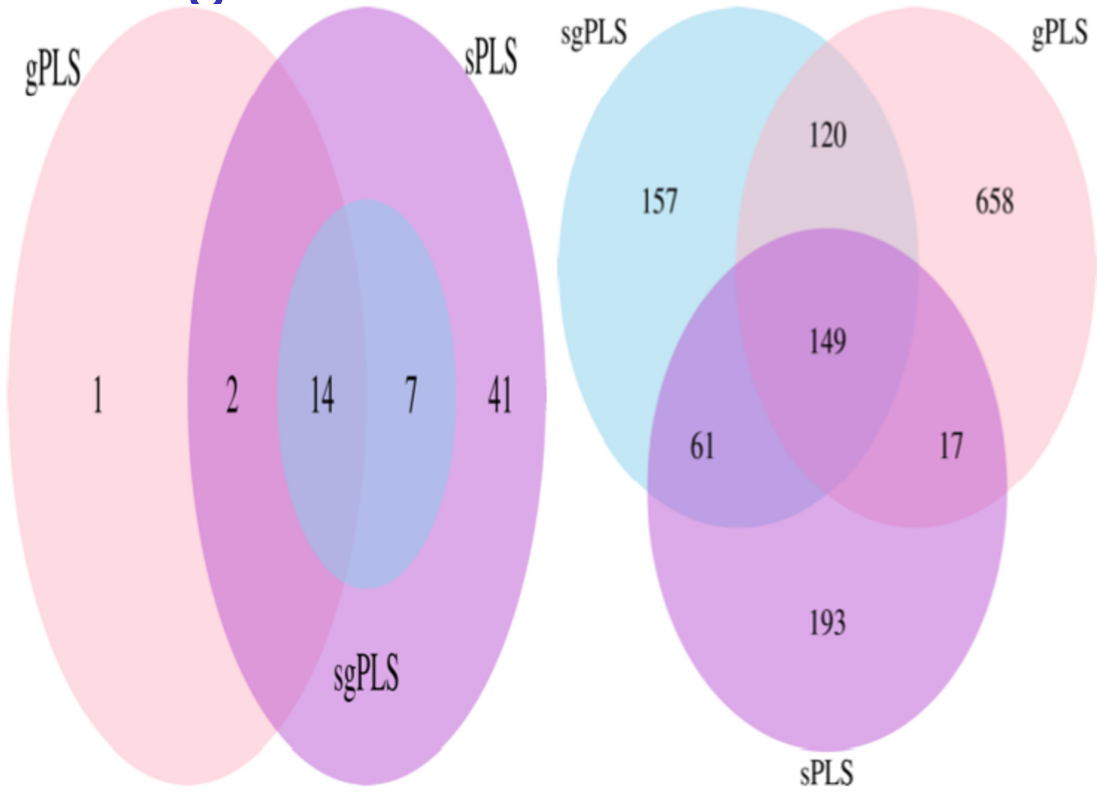
Results: Venn diagram



{

- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)
- Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by the sPLS method.
- sgPLS selected slightly more modules than gPLS (4 more, 14/21 in common). .

Results: Venn diagram



{

- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)
- Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by the sPLS method.
- sgPLS selected slightly more modules than gPLS (4 more, 14/21 in common). .
- However, gPLS led to more genes selected than sgPLS (944)
- In this application, the sgPLS approach led to a parsimonious selection of modules and genes that sound very relevant biologically

sparse group subgroup PLS

Taking into account one more layer in the group structure:

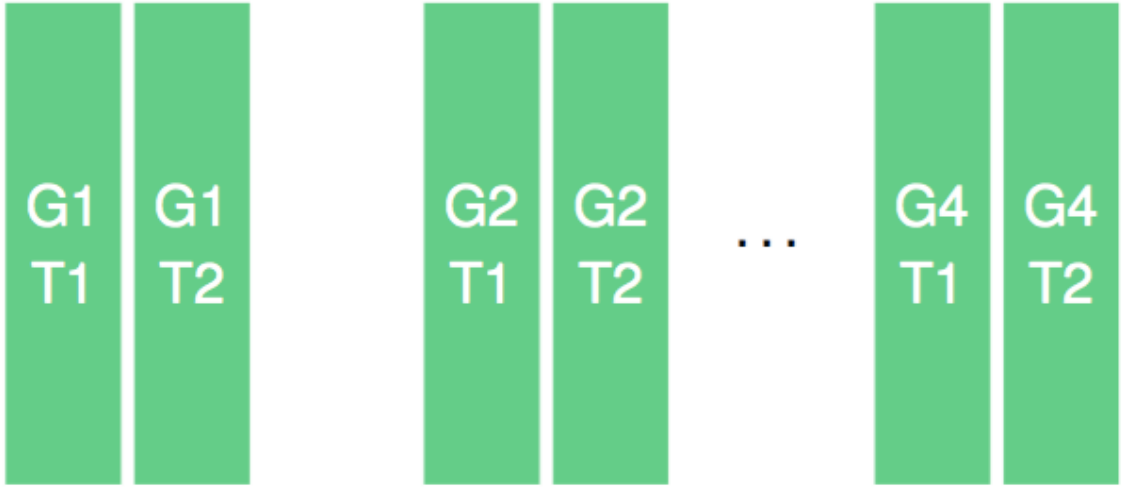
- Example: SNP $\subset$ Gene $\subset$ Pathways
- Longitudinal study

Longitudinal group structures:

- Time index: genes within the same pathway at the same time index have similar functions in regulating a biological system.2

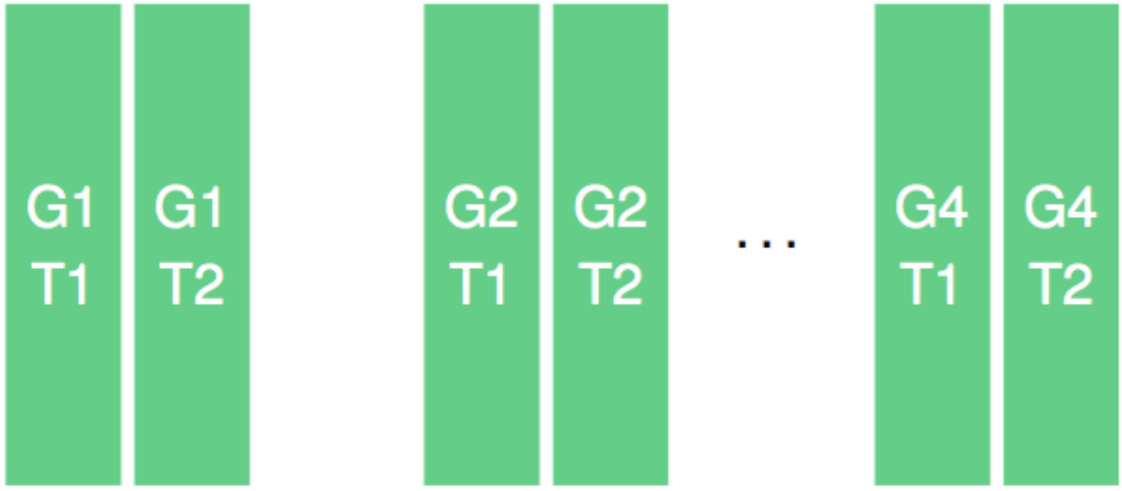
$$\mathbf{G1} = [\underset{G1T1}{\text{gene}_1, \ldots, \text{gene}_k} \mid \underset{G1T2}{\text{gene}_1, \ldots, \text{gene}_k}]$$

Longitudinal group structures:



$$\mathbf{X} = [\underbrace{G1T1, G1T2}_{G1} \mid \underbrace{G2T1, G2T2}_{G2} \mid \cdots \mid \underbrace{G4T1, G4T2}_{G4}]$$

Aims:



- Identify important modules at a group level, important times at a subgroup level and single genes at an individual level.

sparse group subgroup PLS: sgsPLS

$$\xi = \overbrace{\overbrace{0 \times X_1 + 0 \times X_2}^{Time\ 1} + \overbrace{0 \times X_1 + 0 \times X_2}^{Time\ 2}}^{Module\ 1} + \cdots$$

$$+ \overbrace{\overbrace{u_{p-1} \times X_{p-1} + 0 \times X_p}^{Time\ 1} + \overbrace{0 \times X_{p-1} + 0 \times X_p}^{Time\ 2}}^{Module\ k}$$

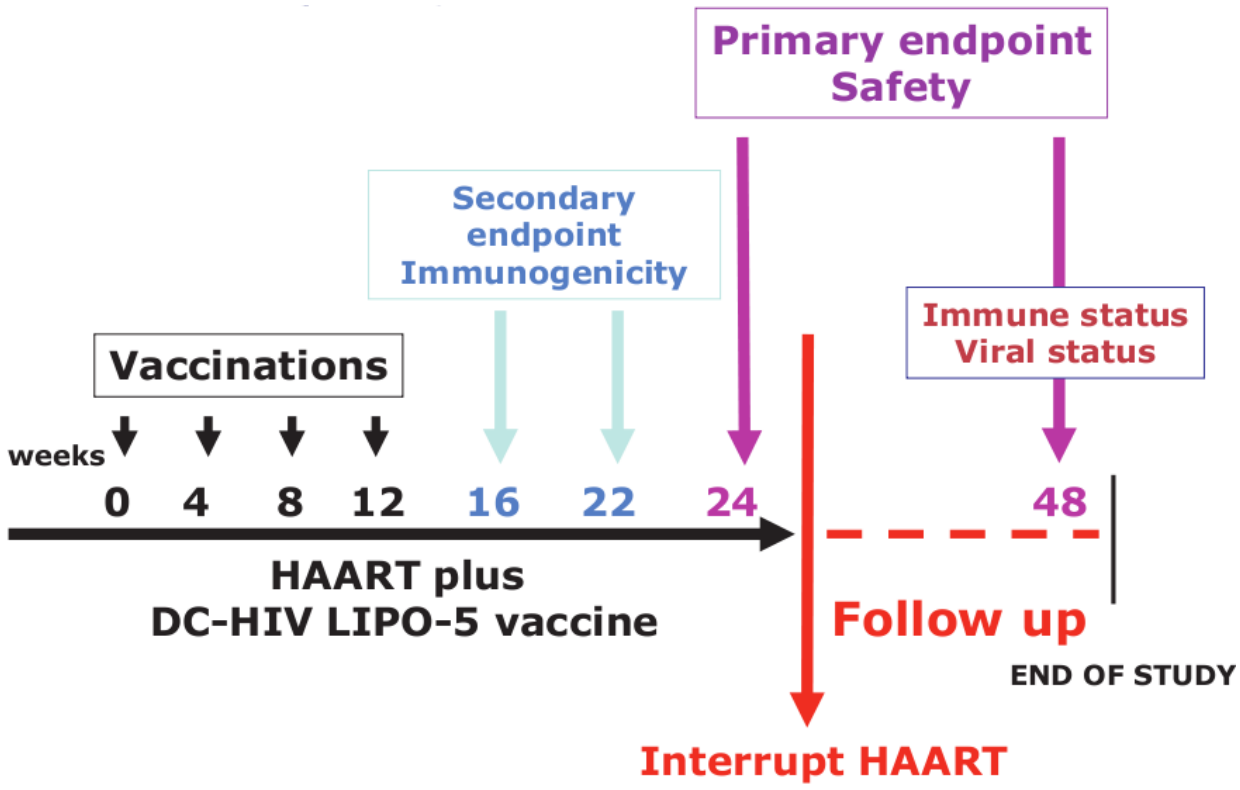
Optimisation of the weights

- X-score $\xi_h = \mathbf{X}_{h-1}\mathbf{u}_h$, Y-score $\omega_h = \mathbf{Y}_{h-1}\mathbf{v}_h$

$$\max_{\mathbf{v}_h,\ \mathbf{u}_h} \ Cov(\mathbf{Xu}, \mathbf{Yv}) - \lambda_1 \sum_{k=1}^{K} ||\mathbf{u}^{(k)}||_2 - \lambda_2 \sum_{k=1}^{K} \sum_{a=1}^{A_k} ||\mathbf{u}^{(k,a)}||_2 - \lambda_3 ||\mathbf{u}||_1$$

such that $\mathbf{v}_h^T \mathbf{v}_h \leqslant 1$ and $\mathbf{u}_h^T \mathbf{u}_h \leqslant 1$.

DALIA application



Data structure

- Significant changes in 69 modules were identified prior to the antiretroviral treatment interruption.
- There are 5399 genes associated to these 69 modules.
- At each of the times Wm4, W0, W4, W8, W12, W16 the gene expressions were measured for the 19 patients.
- At W16, the immune response was evaluated using a set of cytokines.

Data structure

- Response variables **Y** = The immue markers composed of cytokines: IL21, IL2, IL13, IFNg, Luminex score, TH1 score, CD4 ($q = 7$).
- Predictor variables **X** = 5399 gene expressions measured at 4 time points W4, W8, W12, W16 from the 69 extracted modules ($p = 21596$, $n = 19$).

Preliminary results – selected variables

- 19 modules, 784 genes total of 1452 selected variables.

	size.group	consistent	W4	W8	W12	W16
M3.2	126	9	53	42	31	42
M4.1	60	0	24	7	5	7
M4.13	72	3	35	15	16	27
M4.15	41	6	17	11	13	15
M4.2	43	5	15	7	10	14
M4.6	104	5	33	26	26	28
M4.7	82	2	31	15	15	16
M5.1	214	9	36	40	35	47
M5.14	54	3	26	8	8	13
M5.15	24	14	18	18	19	20
M5.5	211	2	77	25	27	30
M5.7	119	3	31	15	13	19
M6.13	38	2	13	8	8	10
M6.14	33	1	7	8	5	8
M6.6	40	2	12	8	17	21
M6.9	35	1	15	5	4	4
M7.1	150	9	33	35	25	41
M7.27	29	1	11	2	3	8
M8.14	27	0	6	4	8	2

R Package

sgPLS available on CRAN

```
library(sgPLS)
example("gPLS")
```

sgsPLS Available now on GITHUB
https://github.com/matt-sutton/sgspls

```
library(devtools)
install_github("matt-sutton/sgspls")
```

Big sgPLS

`bigsgPLS` is an R package that provides an implementation of the two block PLS methods. The method makes use of bigmemory and matrix algebra by chunks to deal with datasets too large for R.

A preliminary paper describing the PLS methods and some of the statistical properties is avaliable on ArXiv Pre-prints https://arxiv.org/abs/1702.07066

```
library(devtools)
install_github("matt-sutton/bigsgPLS",
                host = "https://api.github.com")
```

An example of PLS on the EMNIST dataset is provided here

https://github.com/matt-sutton/bigsgPLS/blob/master/Examples/Example-3-PLS.md

References

PLS

- Wold, H. (1966a) “Nonlinear Estimation by Iterative Least Square Procedures.” In Research Papers in Statistics. *Festschrift for J. Neyman,* edited by F. N. David, 411-444. Wiley.
- Wold, S. (1995) “Chemometrics; what do we mean with it, and what do we want from it?” Chemometrics and Intelligent Laboratory Systems, 30, 109-115.

Extension Sparse PLS

- Lê Cao, K.A., D. Rossouw, C. Robert-GraniÂŽe, and P. Besse (2008) “A sparse PLS for variable selection when integrating omics data” Statistical applications in genetics and molecular biology 7(1):35.
- Chun, H. and S. Keleș (2010) “Sparse partial least squares regression for simultaneous dimension reduction and variable selection.” J R Stat Soc Series B Stat Methodol, 72(1):3-25.
- Liquet B, de Micheaux PL, Hejblum BP, Thiébaut R. (2016) “Group and sparse group partial least square approaches applied in genomics context” Bioinformatics, 32(1):35-42.
- Sutton, Matthew, Rodolphe Thiébaut, and Benoît Liquet. 2018. ”Sparse Partial Least Squares with Group and Subgroup Structure.” Statistics in Medicine 37 (23). Wiley Online Library: 3338–56.
- Lafaye de Micheaux, Pierre, Benoit Liquet, and Matthew Sutton. 2019. ”PLS for Big Data: A unified parallel algorithm for regularised group PLS” Statistics Surveys.

DALIA data

- Lêvy Y, Thiébaut R, Montes M, Lacabaratz C, Sloan L, King B, PÃ©rusat S, Harrod C, Cobb A, Roberts LK, Surenaud M, Boucherie C, Zurawski S, Delaugerre C, Richert L, Chêne G, Banchereau J, Palucka K. (2014) “Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load.” Eur J Immunol. 44(9):2802-10.