Group and Sparse Group Partial Least Square Approaches

Applied in Genomics Context

Benoit Liquet[1,2], Pierre Lafaye de Micheaux[3], Boris Heljbum[4,5], Rodolphe Thiébaut[4,5]

[1] University de Pau et Pays de l'Adour, LMAP.
[2] ARC Centre of Excellence for Mathematical and Statistical Frontiers,
[3] CREST, ENSAI,
[4] Inria, SISTM,
[5] Vaccine Research Institute, Creteil, France.

Contents

Integrative Analysis

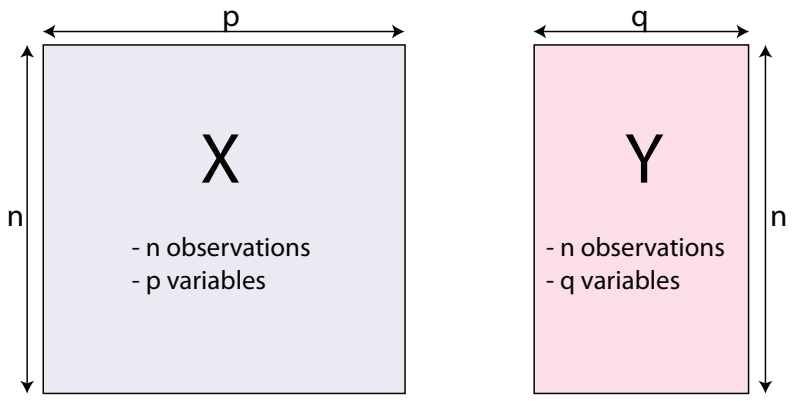
Wikipedia. **Data integration** "involves **combining data** residing in different sources and providing users with a unified view of these data. This process becomes significant in a variety of situations, which include both commercial and **scientific**".

System Biology. **Integrative Analysis:** Analysis of heterogeneous types of data from inter-platform technologies.

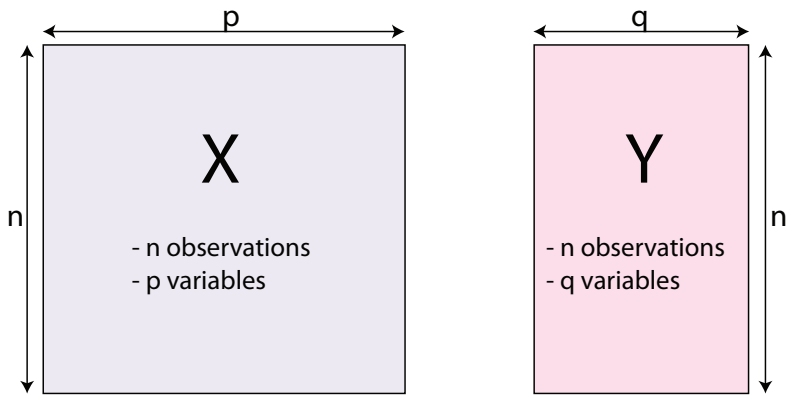
Goal. Combine multiple types of data:

- Contribute to a better understanding of biological mechanism.
- Have the potential to improve the diagnosis and treatments of complex diseases.

Example: Data definition

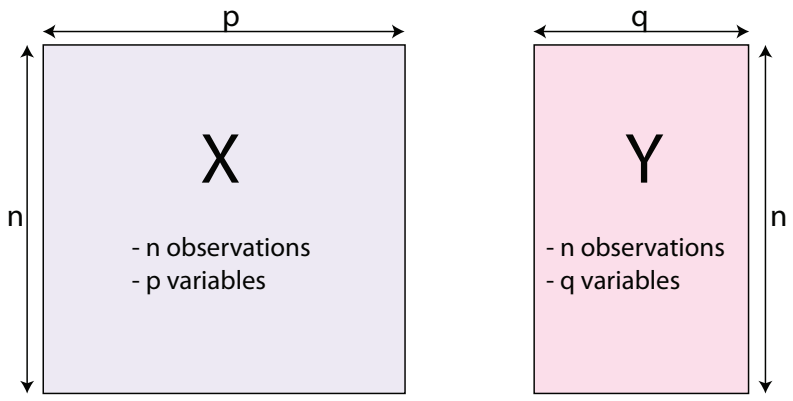


Example: Data definition



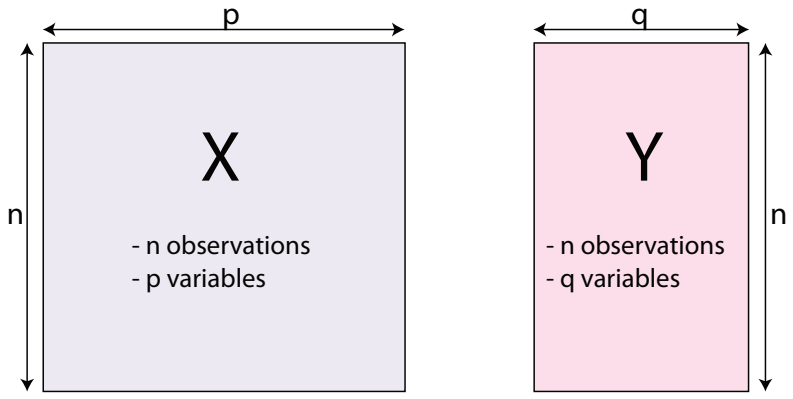
- "Omics." **Y** matrix: gene expression, **X** matrix: SNP (single nucleotide polymorphism). Many others such as proteomic, metabolomic data.

Example: Data definition



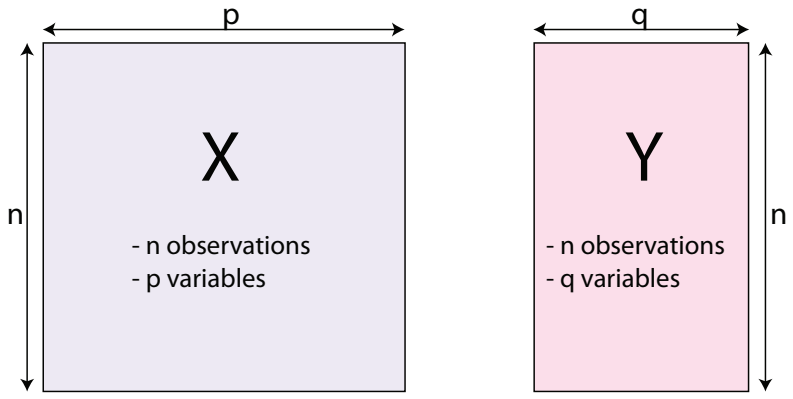
- "Omics." **Y** matrix: gene expression, **X** matrix: SNP (single nucleotide polymorphism). Many others such as proteomic, metabolomic data.
- "neuroimaging". **Y** matrix: behavioral variables, **X** matrix: brain activity (e.g., EEG, fMRI, NIRS)

Example: Data definition



- "Omics." **Y** matrix: gene expression, **X** matrix: SNP (single nucleotide polymorphism). Many others such as proteomic, metabolomic data.
- "neuroimaging". **Y** matrix: behavioral variables, **X** matrix: brain activity (e.g., EEG, fMRI, NIRS)
- "neuroimaging genetics." **Y** matrix: fMRI (Fusion of functional magnetic resonance imaging), **X** matrix: SNP

Example: Data definition



- "Omics." **Y** matrix: gene expression, **X** matrix: SNP (single nucleotide polymorphism). Many others such as proteomic, metabolomic data.
- "neuroimaging". **Y** matrix: behavioral variables, **X** matrix: brain activity (e.g., EEG, fMRI, NIRS)
- "neuroimaging genetics." **Y** matrix: fMRI (Fusion of functional magnetic resonance imaging), **X** matrix: SNP
- "Ecology/Environment." **Y** matrix: Water quality variables , **X** matrix: Landscape variables

Data: Constraints and Aims

- Main constraint: situation with $p > n$

Data: Constraints and Aims

- Main constraint: situation with $p > n$
- Aims:
  1. Symmetric situation. Analysis the associations between two blocks of information, analysis focuses on shared information.

Data: Constraints and Aims

- Main constraint: situation with $p > n$
- Aims:
  1. Symmetric situation. Analysis the associations between two blocks of information, analysis focuses on shared information.
  2. Asymmetric situation. **X** matrix= predictors and **Y** matrix= responses variables, analysis focuses on prediction.

Data: Constraints and Aims

- Main constraint: situation with $p > n$
- Aims:
  1. Symmetric situation. Analysis the associations between two blocks of information, analysis focuses on shared information.
  2. Asymmetric situation. **X** matrix= predictors and **Y** matrix= responses variables, analysis focuses on prediction.

- Partial Least Square Family: dimension reduction approaches

Data: Constraints and Aims

- Main constraint: situation with $p > n$
- Aims:
  1. Symmetric situation. Analysis the associations between two blocks of information, analysis focuses on shared information.
  2. Asymmetric situation. **X** matrix= predictors and **Y** matrix= responses variables, analysis focuses on prediction.

- Partial Least Square Family: dimension reduction approaches
  - PLS find pairs of latent vectors $\mathbf{C_X} = \mathbf{Xu}$, $\mathbf{C_Y} = \mathbf{Yv}$ with maximal covariance.

$$e.g., \quad \mathbf{C_X} = u_1 \times SNP_1 + u_2 \times SNP_2 + \ldots + u_p \times SNP_p$$

  - Symmetric situation and Asymmetric situation.
  - Successive matrix decomposition of **X** and **Y** into new latent variables.

PLS and sparse PLS

PLS

- Output of PLS: K pairs of latent variables $(\mathbf{C_X}^k, \mathbf{C_Y}^k)$, $k = 1, \ldots, K$ with $K << min(p, q)$.
- Reduction method but no variable selection for extracting the most relevant variables from each latent variables.

PLS and sparse PLS

PLS

- Output of PLS: K pairs of latent variables $(\mathbf{C_X}^k, \mathbf{C_Y}^k)$, $k = 1, \ldots, K$ with $K << min(p, q)$.
- Reduction method but no variable selection for extracting the most relevant variables from each latent variables.

sparse PLS

- sparse PLS select the relevant SNPs
- Some coefficients u_l are equal to 0

$$C^k = u_1 \times SNP_1 + \underbrace{u_2}_{=0} \times SNP_2 + \underbrace{u_3}_{=0} \times SNP_3 + \ldots + u_p \times SNP_p$$

- The sPLS components are linear combinations of the selected variables

Group structures within the data

- Natural example: Categorical variables which is a group of dummies variables in a regression setting.

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.
  ↪ These genes can add up to have a larger effect
  ↪ can be detected as a group (i.e., at a pathway or gene set/module level).

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

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

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

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

Group PLS

Aim: Select group variables taking into account the data structures

Group PLS

Aim: Select group variables taking into account the data structures

- PLS components
  $C^k = u_1 \times X_1 + u_2 \times X_2 + u_3 \times X_3 + \ldots + u_p \times X_p$
- sparse PLS components (sPLS)
  $C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$

Group PLS

Aim: Select group variables taking into account the data structures

- PLS components
  $C^k = u_1 \times X_1 + u_2 \times X_2 + u_3 \times X_3 + \ldots + u_p \times X_p$
- sparse PLS components (sPLS)
  $C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$
- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

$\hookrightarrow$ select group of variables; all the variables within a group are selected otherwise none of them are selected

Group PLS

Aim: Select group variables taking into account the data structures

- PLS components
  $C^k = u_1 \times X_1 + u_2 \times X_2 + u_3 \times X_3 + \ldots + u_p \times X_p$
- sparse PLS components (sPLS)
  $C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$
- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

$\hookrightarrow$ select group of variables; all the variables within a group are selected otherwise none of them are selected

does not achieve sparsity within each group

Sparse Group PLS

Aim: combine both sparsity of groups and within each group.
Example, $\boldsymbol{X}$ matrix= genes, we might be interested in identifying particularly important genes in pathways of interest.

- sparse PLS components (sPLS)

$$C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$$

- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

Sparse Group PLS

Aim: combine both sparsity of groups and within each group.
Example, $\boldsymbol{X}$ matrix= genes, we might be interested in identifying particularly important genes in pathways of interest.

- sparse PLS components (sPLS)

$$C^k = u_1 \times X_1 + \underbrace{u_2}_{=0} \times X_2 + \underbrace{u_3}_{=0} \times X_3 + \ldots + u_p \times X_p$$

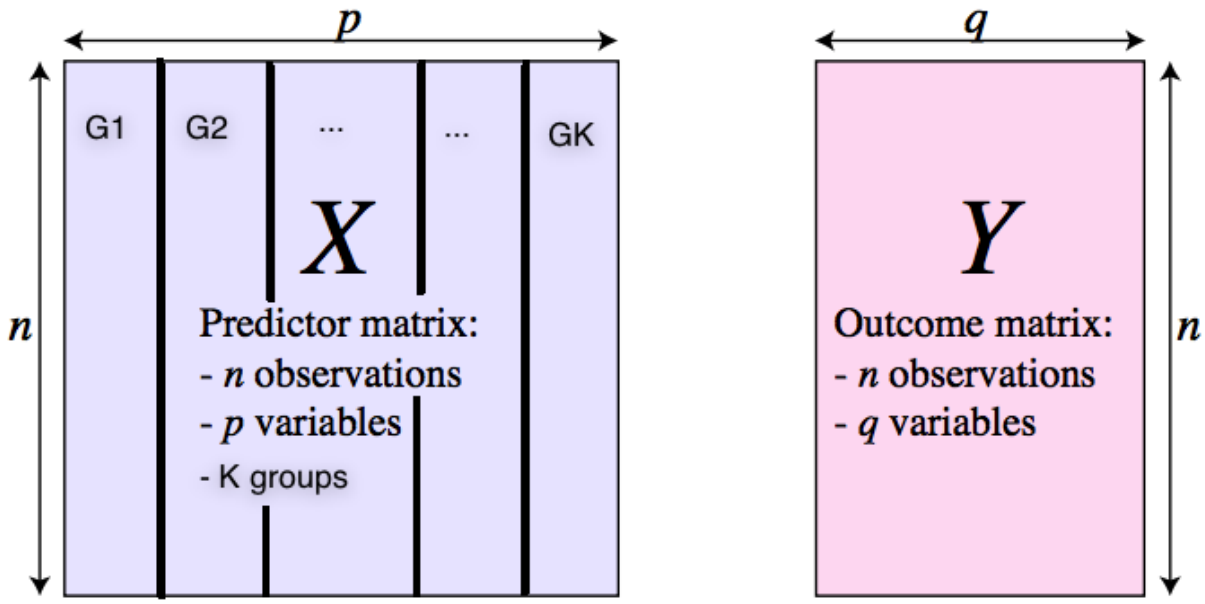
- group PLS components (gPLS)

$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{\neq 0} X_1 + \underbrace{u_5}_{\neq 0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

- sparse group PLS components (sgPLS)

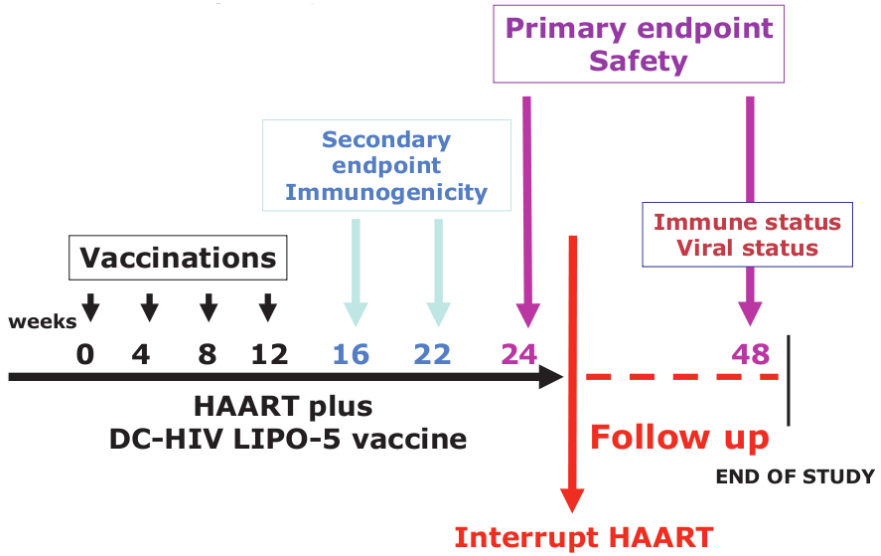
$$C^k = \overbrace{\underbrace{u_1}_{=0} X_1 + \underbrace{u_2}_{=0} X_2}^{module_1} + \overbrace{\underbrace{u_3}_{\neq 0} X_3 + \underbrace{u_4}_{=0} X_4 + \underbrace{u_5}_{=0} X_5}^{module_2} \ldots \overbrace{\underbrace{u_{p-1}}_{=0} X_{p-1} + \underbrace{u_p}_{=0} X_p}^{module_K}$$

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

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 ?

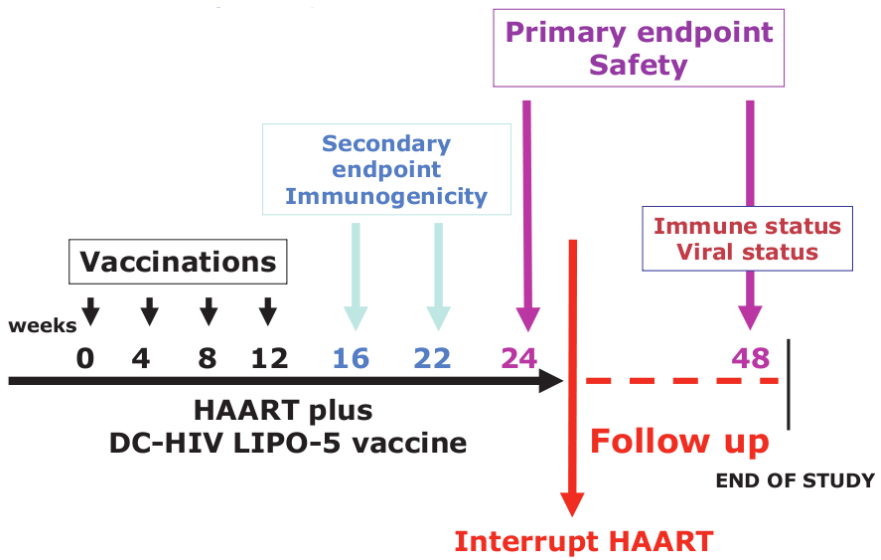
First results obtained using group of genes

- 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:

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

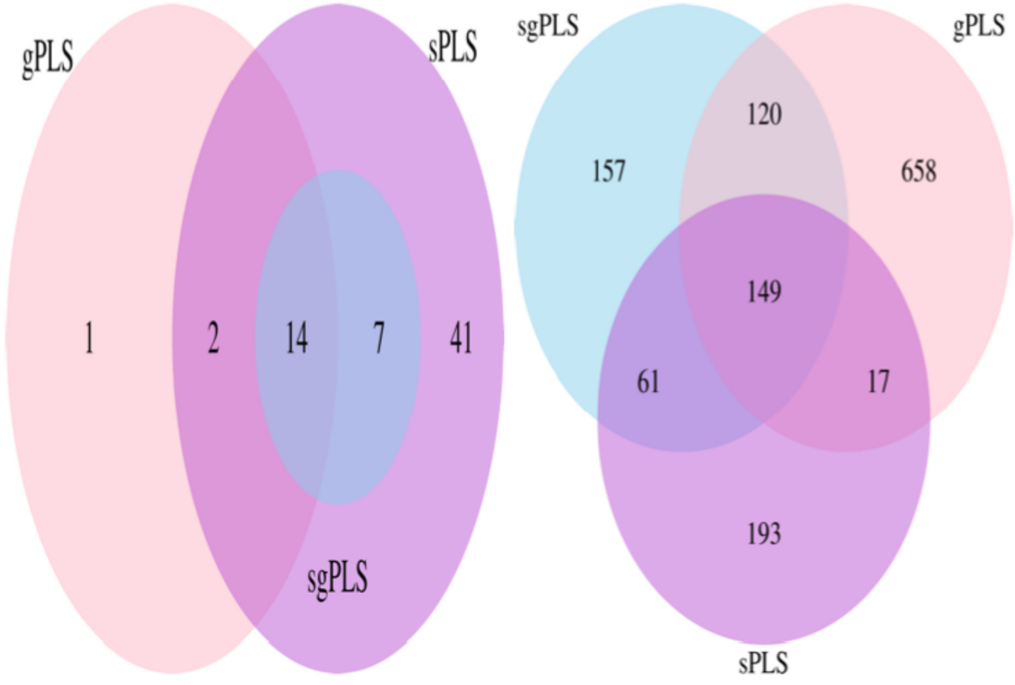
Results: Modules and number of genes selected

		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

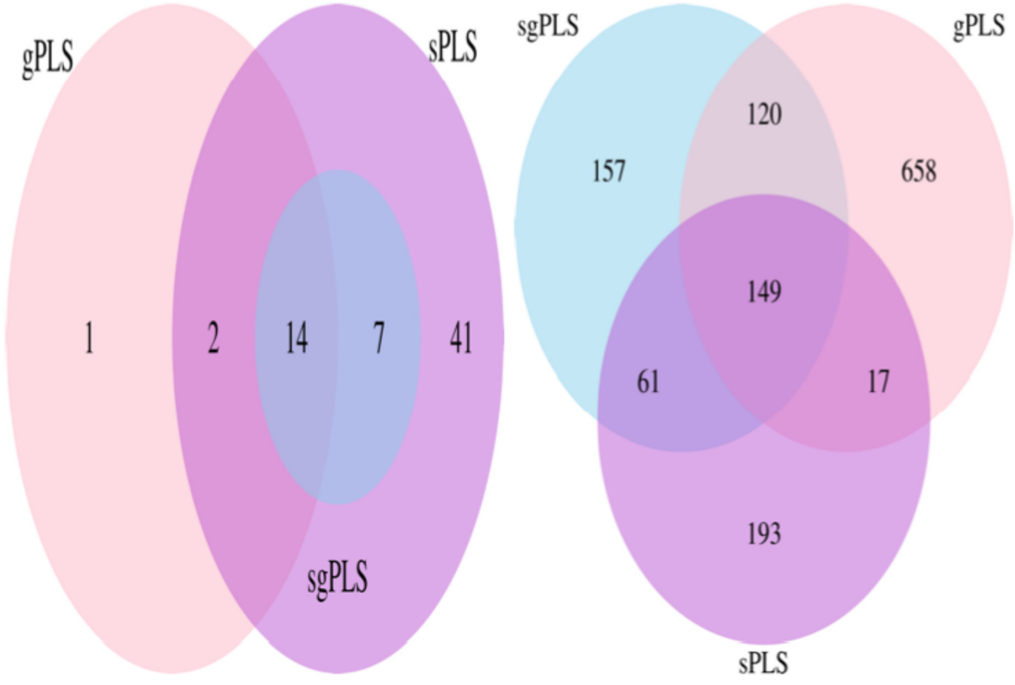
Results: Modules and number of genes selected

		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

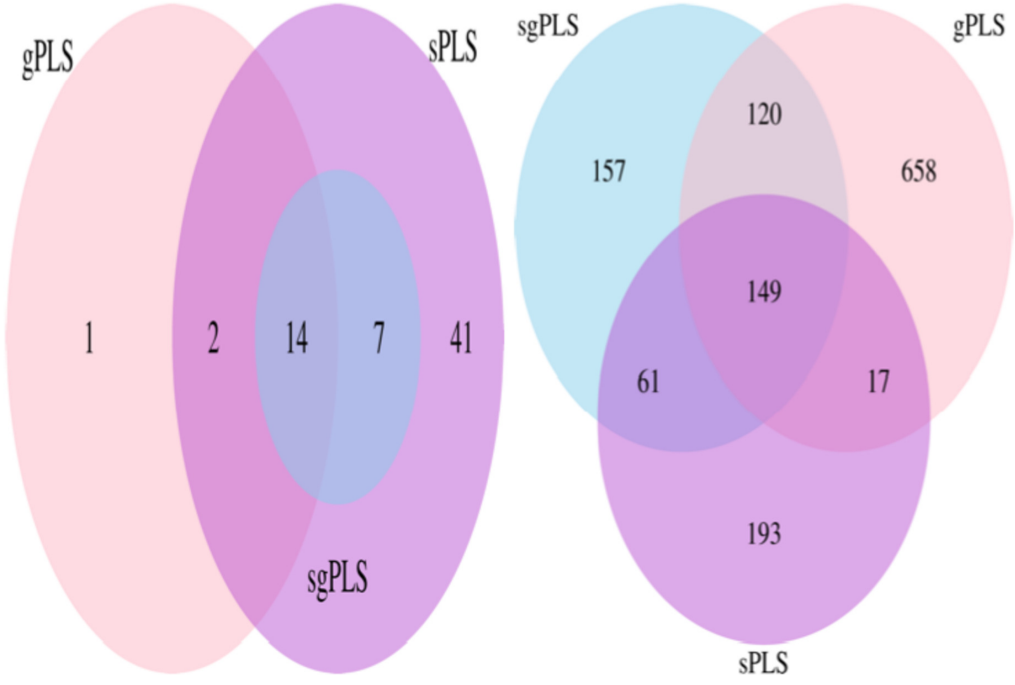


Results: Venn diagram



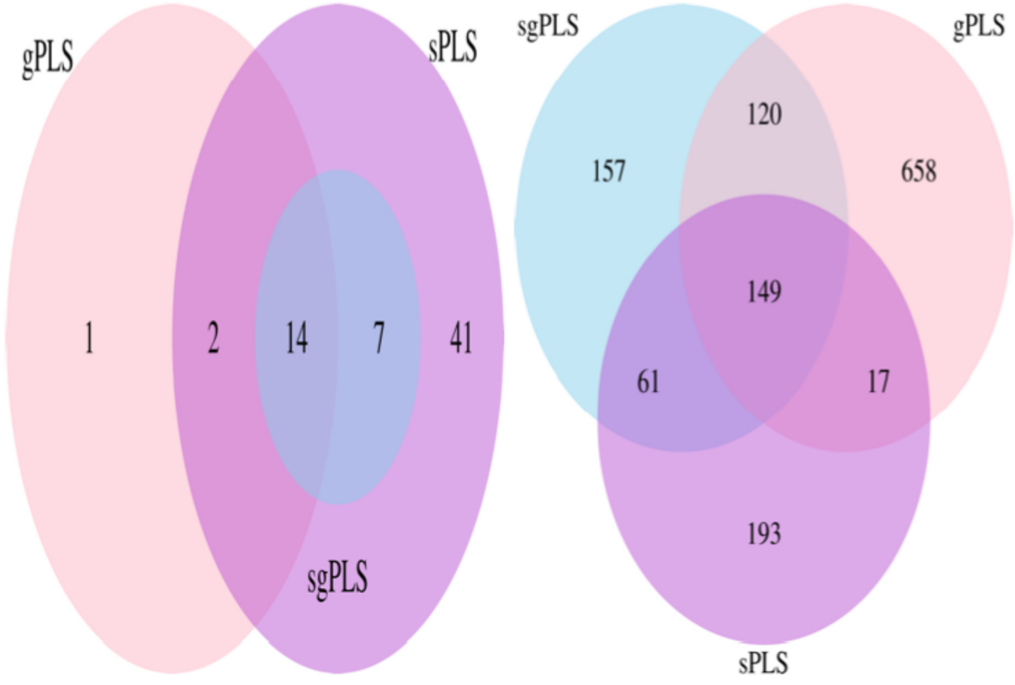
- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)

Results: Venn diagram



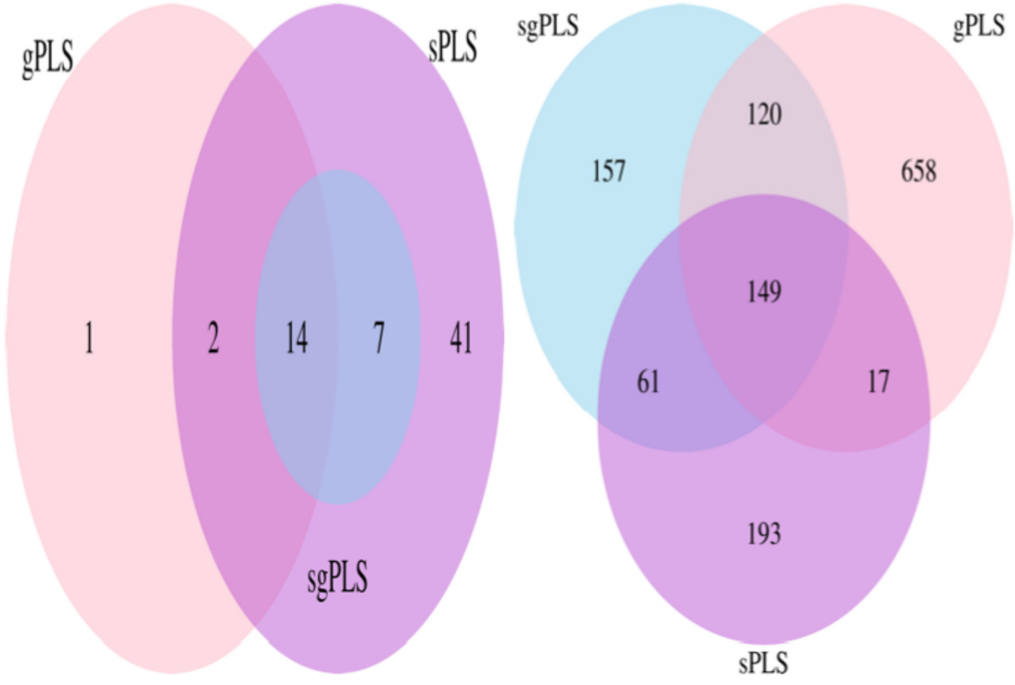
- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)

Results: Venn diagram



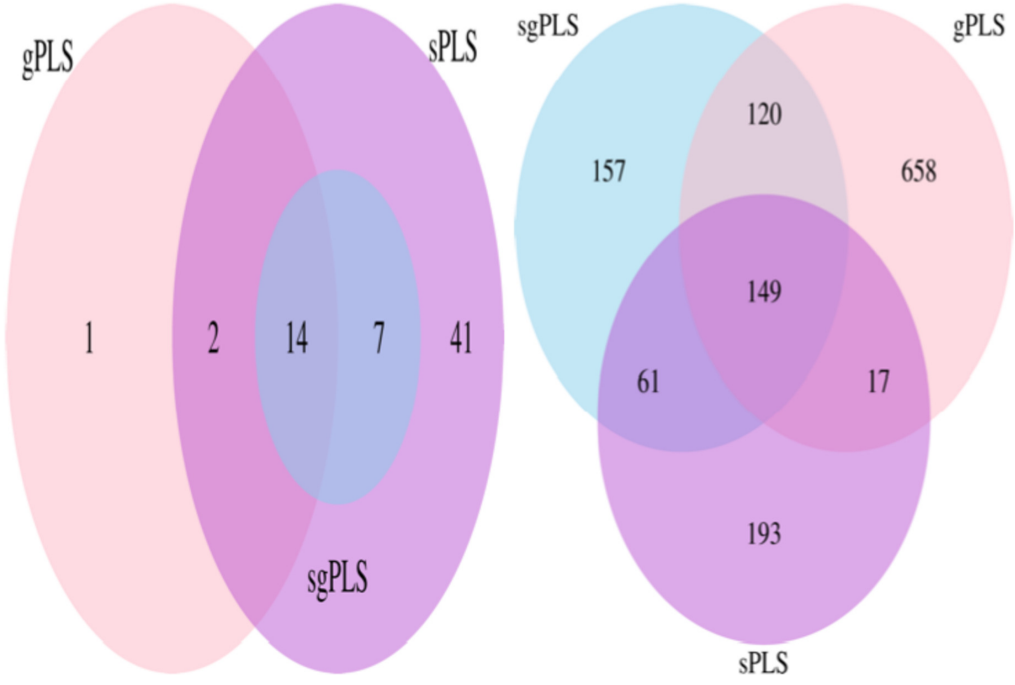
- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)
- Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by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

Results: Venn diagram



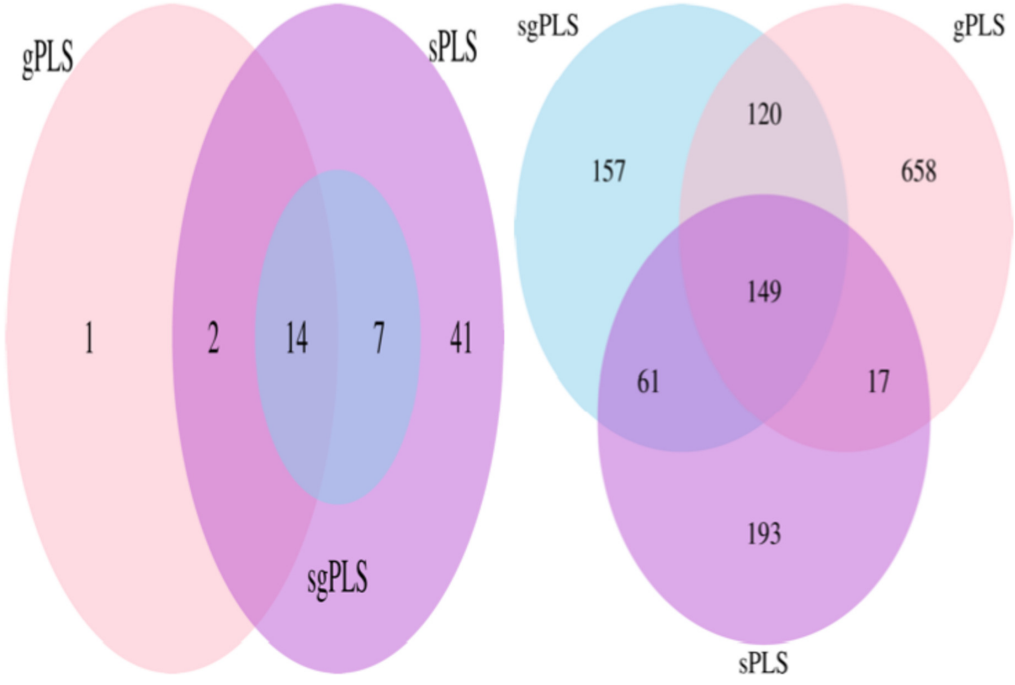
- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)
- Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by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

Results: Venn diagram



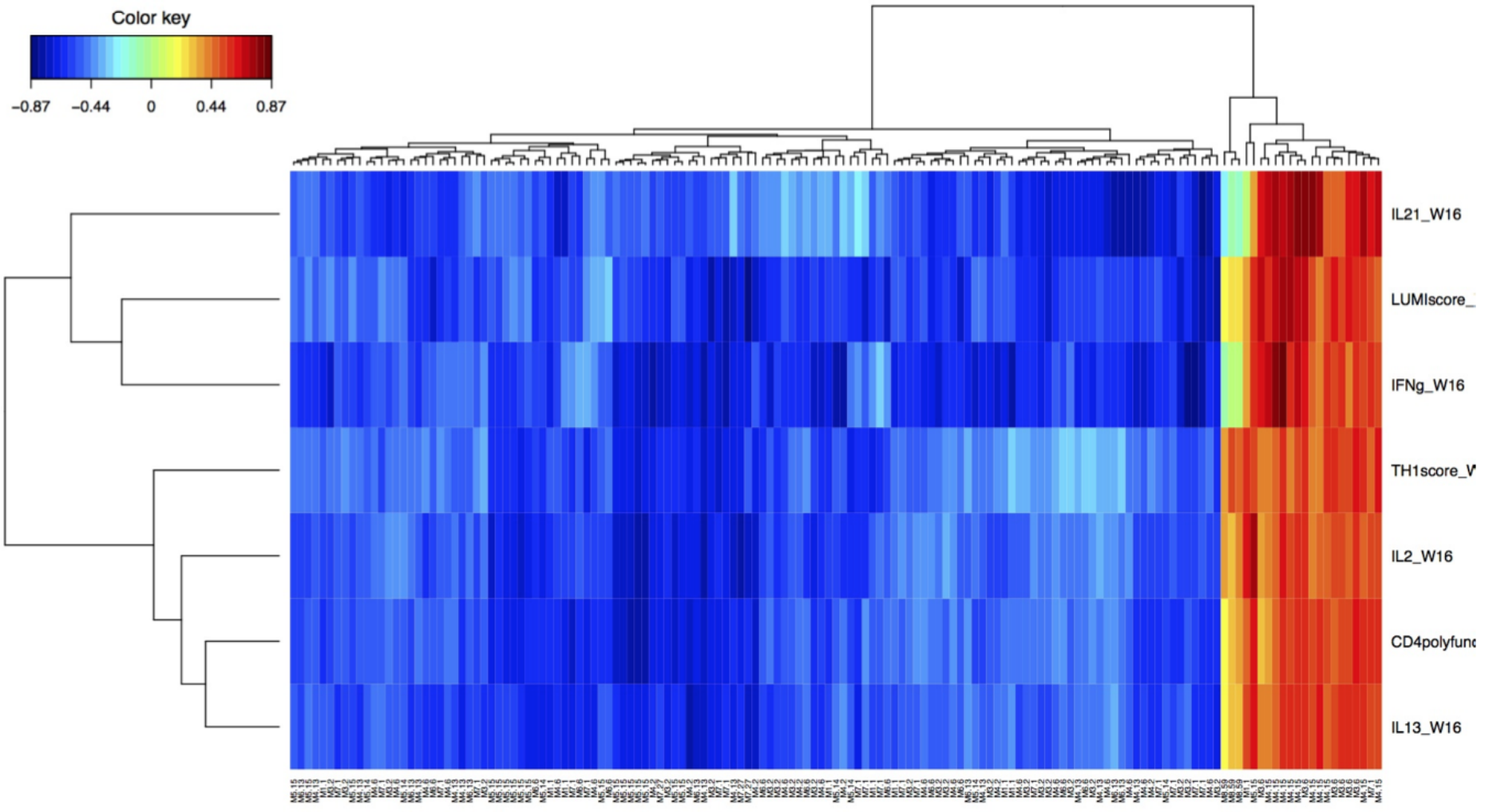
- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)
- Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by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

Results: Venn diagram

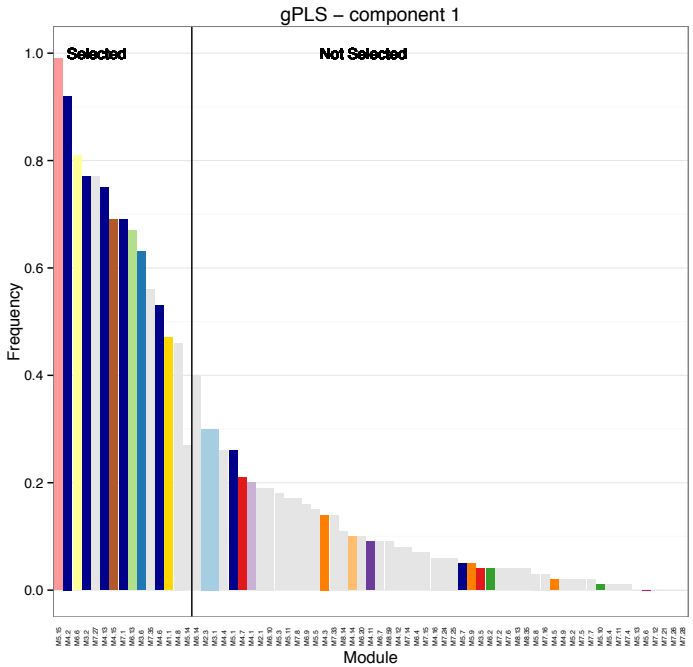


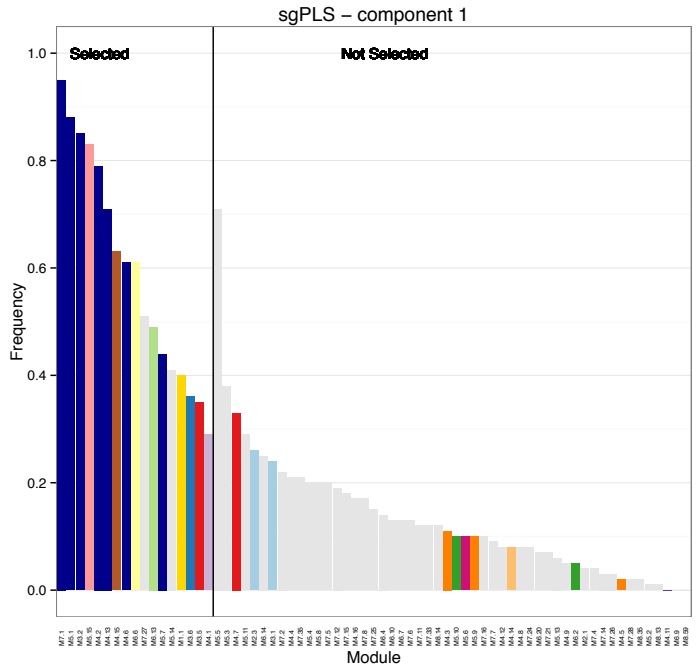
- sgPLS methods selected slightly more genes than the sPLS (respectively 487 and 420 genes selected)
- But sgPLS selected fewer modules than the sPLS (respectively 21 and 64 groups of genes selected by sPLS)
- Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by 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 Chaussabel's functional modules: http://www.biir.net/public_wikis/module_annotation/V2_Trial_8_Modules

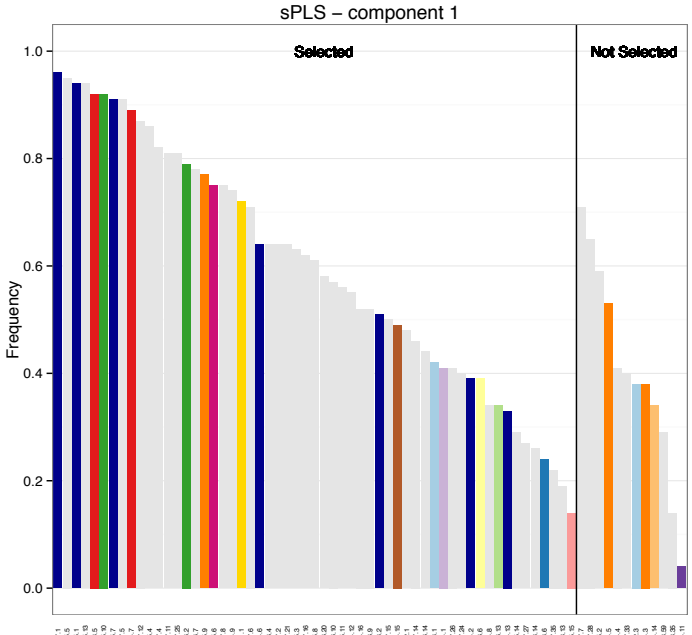
Visualisation of these associations



Stability of the variable selection (100 bootstrap samples)







Apoptosis / Survival

- Apoptosis / Survival
- Cell Cycle
- Cell Death
- Cytotoxic/NK Cell
- Erythrocytes
- Inflammation
- Mitochondrial Respiration
- Mitochondrial Stress / Proteasome
- Monocytes
- Neutrophils
- Plasma Cells
- Platelets
- Protein Synthesis
- T cell
- T cells
- Undetermined

Now some mathematics ...

PLS family

PLS: Partial Least Squares or Projection to Latent Structures

(i) Partial Least Squares Correlation (PLSC) also called PLS-SVD,

(ii) PLS in mode A (PLS-W2A, for Wold's Two-Block, Mode A PLS),

(iii) PLS in mode B (PLS-W2B) also called Canonical Correlation Analysis (CCA)

(iv) Partial Least Squares Regression (PLSR, or PLS2).

PLS family

PLS: Partial Least Squares or Projection to Latent Structures

(i) Partial Least Squares Correlation (PLSC) also called PLS-SVD,

(ii) PLS in mode A (PLS-W2A, for Wold's Two-Block, Mode A PLS),

(iii) PLS in mode B (PLS-W2B) also called Canonical Correlation Analysis (CCA)

(iv) Partial Least Squares Regression (PLSR, or PLS2).

- (i),(ii) and (iii) are symmetric while (iv) is asymmetric.
- Different objective functions to optimise.
- Good news: all are based on the singular value decomposition (SVD).

Singular Value Decomposition (SVD)

Definition 1

Let a matrix $\mathcal{M} : p \times q$ of rank r:

$$\mathcal{M} = \mathcal{U}\Delta\mathcal{V}^{\mathsf{T}} = \sum_{l=1}^{r} \delta_l \boldsymbol{u}_l \boldsymbol{v}_l^{\mathsf{T}}, \tag{1}$$

- $\mathcal{U} = (\boldsymbol{u}_l) : p \times r$ and $\mathcal{V} = (\boldsymbol{v}_l) : q \times r$ are two orthogonal matrices which contain the normalised left (resp. right) singular vectors
- $\Delta = \mathrm{diag}(\delta_1, \ldots, \delta_r)$: the ordered singular values $\delta_1 \geqslant \delta_2 \geqslant \cdots \geqslant \delta_r$.

Connexion between SVD and maximum covariance

Optimization problem of the PLS:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}{\operatorname{argmax}} Cov(\mathbf{X}\boldsymbol{u}, \mathbf{Y}\boldsymbol{v}), \qquad h = 1, \ldots, r,$$

Connexion between SVD and maximum covariance

Optimization problem of the PLS:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}{\operatorname{argmax}} Cov(\mathbf{X}\boldsymbol{u}, \mathbf{Y}\boldsymbol{v}), \qquad h = 1, \ldots, r,$$

The solution is given by the SVD of $\boldsymbol{\mathcal{M}} = \mathbf{X}^\mathsf{T}\mathbf{Y}$:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = (\boldsymbol{u}_1, \boldsymbol{v}_1)$$

Connexion between SVD and maximum covariance

Optimization problem of the PLS:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1}{\operatorname{argmax}} Cov(\mathbf{X}\boldsymbol{u}, \mathbf{Y}\boldsymbol{v}), \qquad h = 1, \ldots, r,$$

The solution is given by the SVD of $\boldsymbol{\mathcal{M}} = \mathbf{X}^\mathsf{T}\mathbf{Y}$:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = (\boldsymbol{u}_1, \boldsymbol{v}_1)$$

Why is it useful ?

SVD properties

Theorem 2

Eckart-Young (1936) states that the SVD provides the best reconstitution (in a least squares sense) of a given matrix $\boldsymbol{\mathcal{M}}$ by a matrix with a lower rank:

$$\min_{\mathcal{A} \text{ of rank } k} \|\boldsymbol{\mathcal{M}} - \mathcal{A}\|_F^2 = \sum_{l=k+1}^{r} \delta_l^2 = \left\| \boldsymbol{\mathcal{M}} - \sum_{l=1}^{k} \delta_l \boldsymbol{u}_l \boldsymbol{v}_l^{\mathsf{T}} \right\|_F^2 .$$

If the minimum is searched for matrices $\mathcal{A}$ of rank 1, which are under the form $\widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^{\mathsf{T}}$ where $\widetilde{\boldsymbol{u}}$, $\widetilde{\boldsymbol{v}}$ are non-zero vectors, we obtain

$$\min_{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}} \left\| \boldsymbol{\mathcal{M}} - \widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^{\mathsf{T}} \right\|_F^2 = \sum_{l=2}^{r} \delta_l^2 = \left\| \boldsymbol{\mathcal{M}} - \delta_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathsf{T}} \right\|_F^2 .$$

SVD properties

Thus, solving

$$\underset{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}}{\operatorname{argmin}} \left\| \boldsymbol{\mathcal{M}} - \widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^{\mathsf{T}} \right\|_F^2 \tag{2}$$

and norming the resulting vectors gives us $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$.

SVD properties

Thus, solving

$$\underset{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}}{\operatorname{argmin}} \left\| \boldsymbol{\mathcal{M}} - \widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^{\mathsf{T}} \right\|_F^2 \tag{2}$$

and norming the resulting vectors gives us $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$.

- Shen and Huang (2008) connected (2) to least square minimisation in regression
  $\hookrightarrow$ rendering possible the use of many existing variable selection techniques using regularisation penalties.

SVD properties

Thus, solving

$$\underset{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}}{\operatorname{argmin}} \left\| \boldsymbol{\mathcal{M}} - \widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^{\mathsf{T}} \right\|_F^2 \tag{2}$$

and norming the resulting vectors gives us $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$.

- Shen and Huang (2008) connected (2) to least square minimisation in regression
  $\hookrightarrow$ rendering possible the use of many existing variable selection techniques using regularisation penalties.
- Same spirit, we propose iterative algorithms to find normed vectors $\widetilde{\boldsymbol{u}}$ and $\widetilde{\boldsymbol{v}}$ that minimise the following penalised sum-of-squares criterion

  $$\left\| \boldsymbol{\mathcal{M}} - \widetilde{\boldsymbol{u}}\widetilde{\boldsymbol{v}}^{\mathsf{T}} \right\|_F^2 + P_\lambda(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}}),$$

  for specific cases of matrix $\boldsymbol{\mathcal{M}}$ and several penalisation terms $P_\lambda(\widetilde{\boldsymbol{u}}, \widetilde{\boldsymbol{v}})$.

$\hookrightarrow$ many sparse versions of the four methods (i)–(iv).

Now some R code ....

Package related to PLS model

- `plsdepot`: contains different methods for PLS analysis of one or two data tables such as Tucker's Inter-Battery, NIPALS, SIMPLS, SIMPLS-CA, PLS Regression, and PLS Canonical Analysis.
- `pls`: Multivariate regression methods Partial Least Squares Regression (PLSR), Principal Component Regression (PCR) and Canonical Powered Partial Least Squares (CPPLS).
- `plspm`: Tools for Partial Least Squares Path Modeling (PLS-PM)
- `spls`: This package provides functions for fitting a Sparse Partial Least Squares Regression and Classification
- `mixOmics`: Omics Data Integration Project including generalised Canonical Correlation Analysis, sparse Partial Least Squares and sparse Discriminant Analysis
- `PMA`: Performs Penalized Multivariate Analysis: a penalized matrix decomposition, sparse principal components analysis, and sparse canonical correlation analysis

Main Packages related to lasso model: univariate response variable

- `glmnet`: Lasso and Elastic-Net Regularized Generalized Linear Models
- `lars`: Least Angle Regression, Lasso and Forward Stagewise
- `penalized`: L1 (Lasso and Fused Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model
- `SGL`: SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization
- `lassoscore`: High-Dimensional Inference with the Penalized Score Test

Main Packages related to lasso model: Multivariate response variable

- `glmnet`: Lasso for multivariate response based on a group penalty
- `MSGLasso`: Multivariate Sparse Group Lasso for computing the multivariate sparse group lasso with complex group structures.

R package: sgPLS

- `sgPLS` package implements sPLS, gPLS and sgPLS methods: `http://cran.r-project.org/web/packages/sgPLS/index.html`
- Including some functions for choosing the tuning parameters related to predictor matrix for different sparse PLS model (regression mode).
- Some simple code to perform a sgPLS method.

```
model.sgPLS <- sgPLS(X, Y, ncomp = 2, mode = "regression",
                     keepX = c(4, 4), keepY = c(4, 4),
                     ind.block.x = ind.block.x ,
                     ind.block.y = ind.block.y,
                     alpha.x = c(0.5, 0.5),
                     alpha.y = c(0.5, 0.5))
```

- Last version includes sparse group Discriminant Analysis.
- Package compatible with many `mixOmics` functions

Concluding Remarks

- Provide two sparse PLS approaches taking into account the data structure
  - group PLS which enables to select group of variables.
  - sparse group PLS which adds some sparsity within group.
- Methods available for the 4 cases of PLS models.
- Simulation and application highlight the advantages of the group PLS and sparse group compared to sparse PLS.
- Methods available through `sgPLS` R package.

Concluding Remarks

- Provide two sparse PLS approaches taking into account the data structure
  - group PLS which enables to select group of variables.
  - sparse group PLS which adds some sparsity within group.
- Methods available for the 4 cases of PLS models.
- Simulation and application highlight the advantages of the group PLS and sparse group compared to sparse PLS.
- Methods available through `sgPLS` R package.
- Extension to other penalty functions:
  - In linear model setting: Garcia et al (2014) proposed method to select important regressor groups, subgroups and individuals.
  - One more layout than the sparse group Lasso.

References

- Yuan,M. and Lin,Y. (2006) *Model selection and estimation in regression with grouped variables*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (1), 49-67.
- Simon,N., Friedman,J., Hastie,T. and Tibshirani,R. (2013) *A sparse-group lasso.* Journal of Computational and Graphical Statistics, 22 (2), 231-245.
- Le Cao,K.A., Rossouw,D., Robert-Granie,C. and Besse,P. (2008) *Sparse PLS: Variable Selection when Integrating Omics data.* Statistical Application and Molecular Biology, 7 (1):37.
- Lin,D., Zhang,J., Li,J., Calhoun,V., Deng,H.W. and Wang,Y.P. (2013) *Group sparse canonical correlation analysis for genomic data integration*. BMC Bioinformatics, 14 (1), 245.
- Garcia,T.P., Muller,S., Carroll,R.J. and Walzem,R.L. (2014) *Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data*. Bioinformatics, 30 (6), 831-837.
- Liquet B., Lafaye de Micheaux P, Hejblum B., Thiebaut R., *Group and Sparse Group Partial Least Square Approaches Applied in Genomics Context.* Bioinformatics, (2016).

ANY QUESTIONS ?