Multiblock supervised analyses

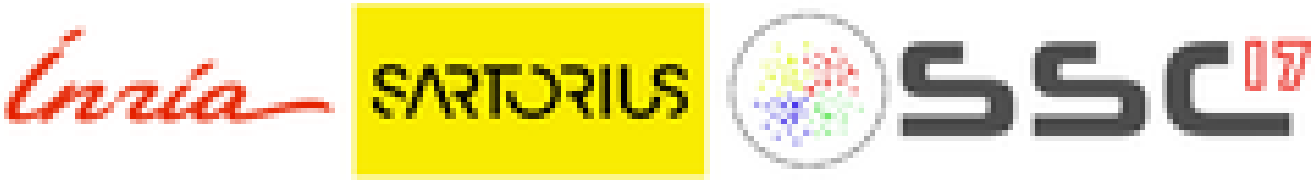
Should we really normalize blocks?

Hadrien Lorenzo[1], Rodolphe Thiébaut[2], Jérôme Saracco[1], Olivier Cloarec[3]

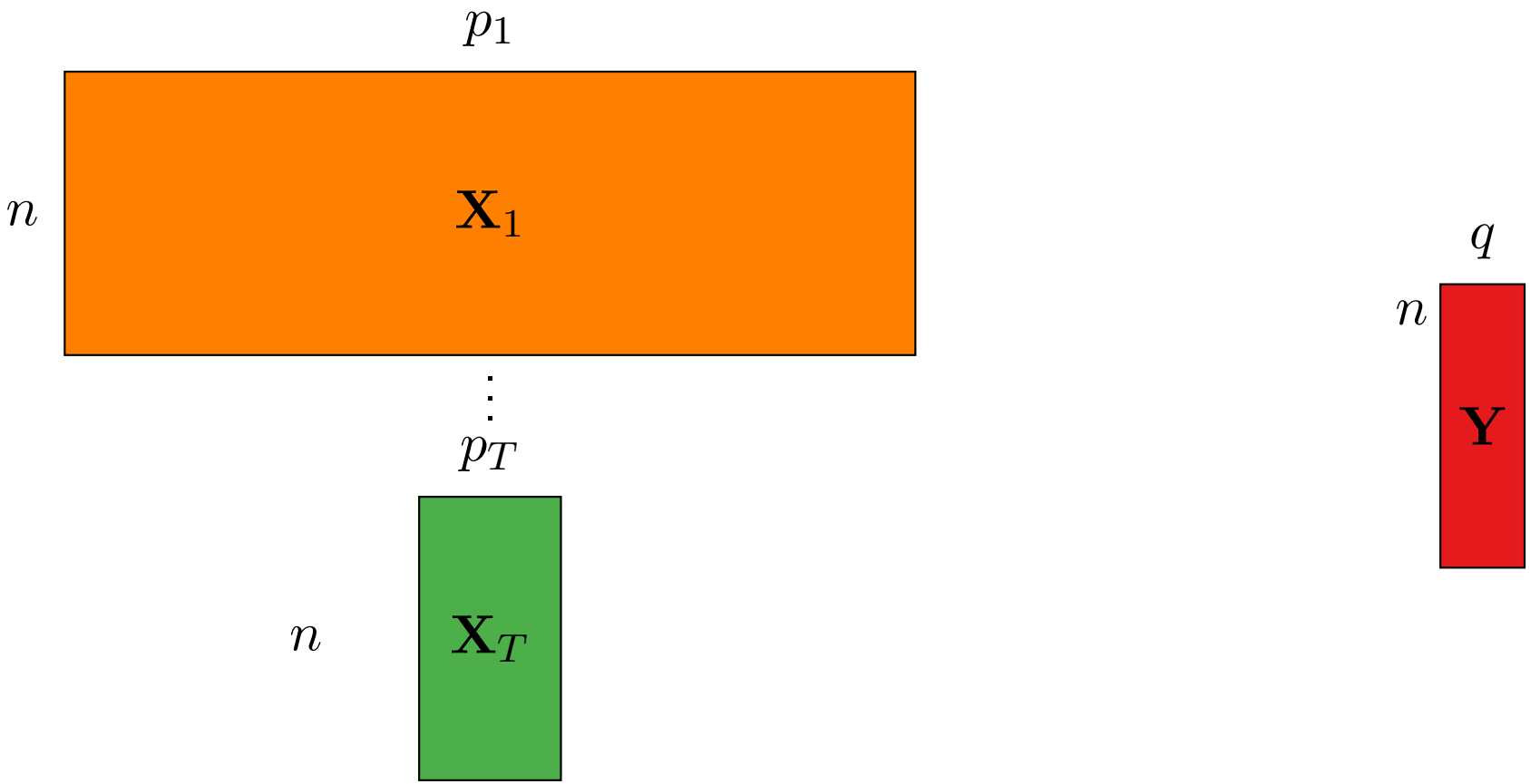
[1]ASTRAL, INRIA [2]SISTM, INRIA

[3]Corporate Research Advanced Data Analytics, Sartorius.

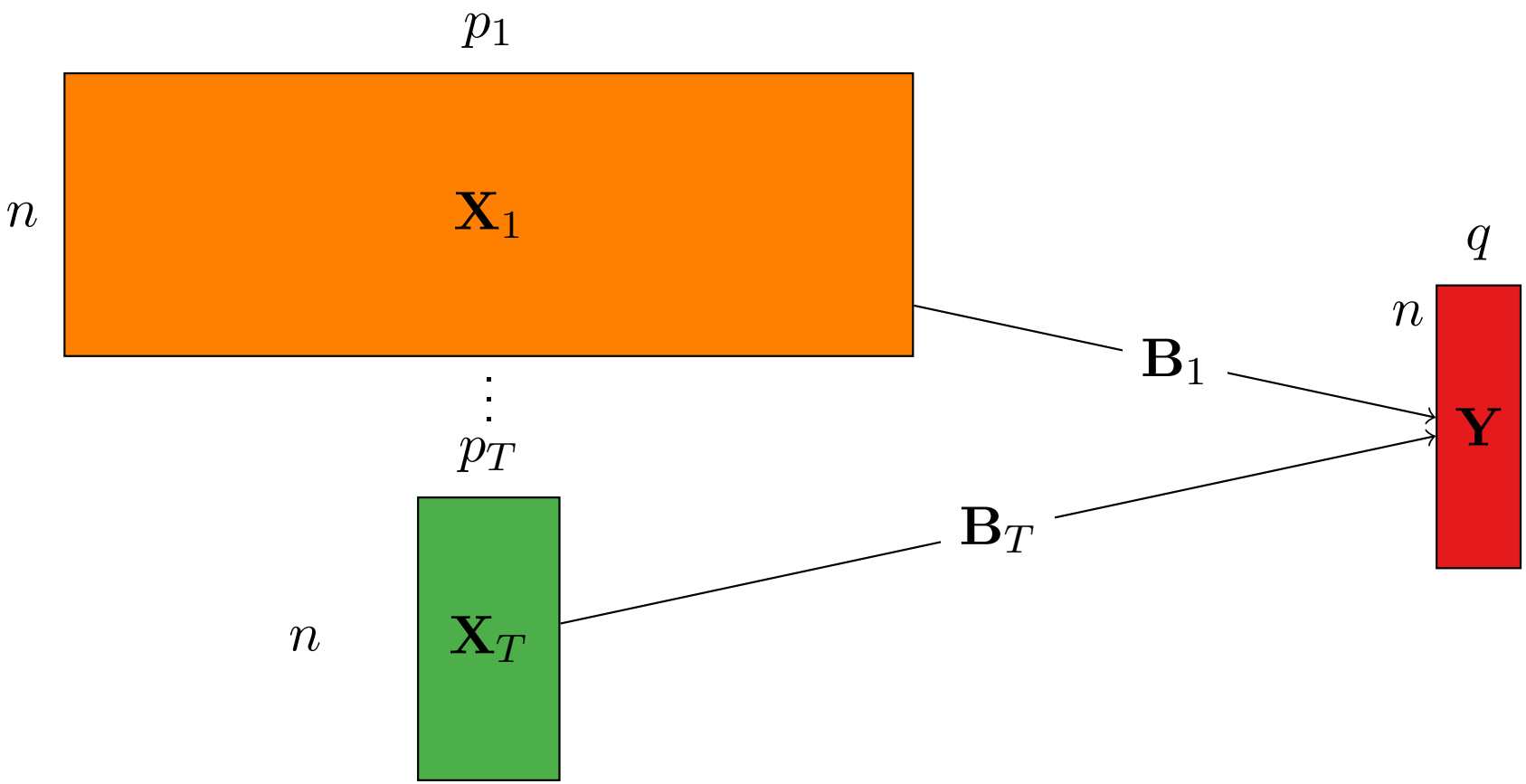
SSC17, Aalborg, Denmark, September 8 2021



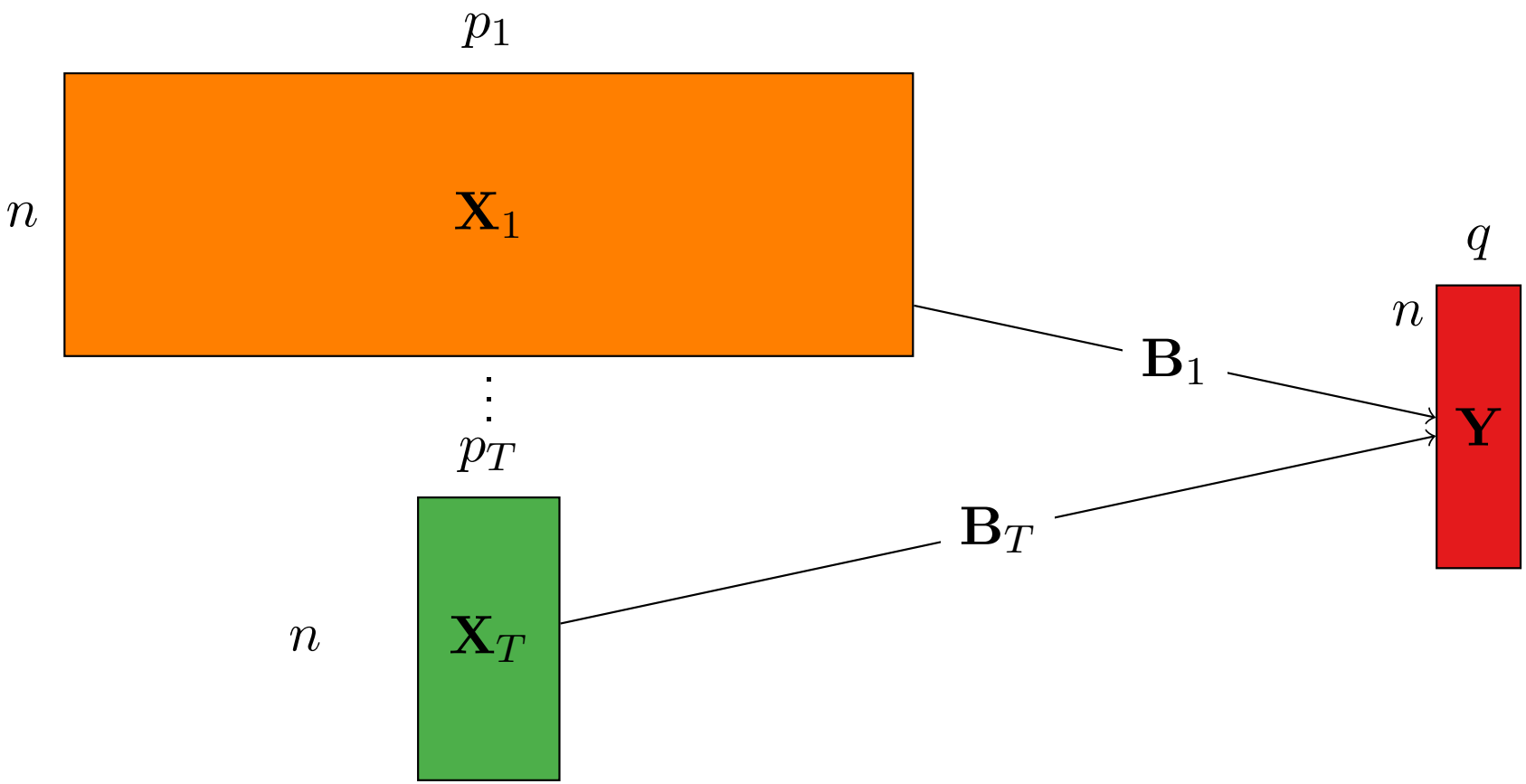
Supervized Multiblock Analyses in linear context



Supervised Multiblock Analyses in linear context



Supervized Multiblock Analyses in linear context



$$\widehat{\mathbf{Y}} = \widehat{\mathbf{B}}_1\mathbf{X}_1 + \cdots + \widehat{\mathbf{B}}_T\mathbf{X}_T$$

Adapt from monoblock supervized analyses

From now on, application to **PLS** Wold (1966) based methodologies.

Adapt classical mono-block analysis such as $\widehat{\mathbf{Y}} = \widehat{\mathbf{B}}\mathbf{X}$.

Different solutions:

- (**V0**) Westerhuis *et al.* (1998): $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_T]$.
- (**V1**) Wold *et al.* (1996): $\mathbf{X} = \left[\mathbf{X}_1/\sqrt{p_1}, \ldots, \mathbf{X}_T/\sqrt{p_T}\right]$.
- (**V2**): $\mathbf{X} = [\mathbf{X}_1/\|\mathbf{X}_1\|, \ldots, \mathbf{X}_T/\|\mathbf{X}_T\|]$.

Remark on the Block Normalization solutions (V1) and (V2)

If variables are standardized, $\forall t = 1 \ldots T,\ \|\mathbf{X}_t\|^2 = np_t$

$$\Longrightarrow \quad \textbf{(V1)}=\textbf{(V2)}$$

Idea and drawbacks of block normalization

The idea

Give the same level of confidence to each block, whatever the size of this block.

Idea and drawbacks of block normalization

The idea

Give the same level of confidence to each block, whatever the size of this block.

The chosen solution

Accept the same proportion of block variance from each block: equivalent SNRs (signal to noise ratios).

Idea and drawbacks of block normalization

A problem for correlated blocks

If the interesting variance proportions from blocks are different:

- $\mathbf{X}_1$ with $p_1 = 20\ 000$ and only 5% of the variance is associated with $\mathbf{Y}$ (Transciptomics, Spectroscopics,...).

Idea and drawbacks of block normalization

A problem for correlated blocks

If the interesting variance proportions from blocks are different:

- $\mathbf{X}_1$ with $p_1 = 20\ 000$ and only 5% of the variance is associated with $\mathbf{Y}$ (Transciptomics, Spectroscopics,...).
- $\mathbf{X}_2$ with $p_2 = 200$ and 20% of the variance is associated with $\mathbf{Y}$ (Metabolomics for example).

Idea and drawbacks of block normalization

A problem for correlated blocks

If the interesting variance proportions from blocks are different:

- $\mathbf{X}_1$ with $p_1 = 20\ 000$ and only 5% of the variance is associated with $\mathbf{Y}$ (Transciptomics, Spectroscopics,...).
- $\mathbf{X}_2$ with $p_2 = 200$ and 20% of the variance is associated with $\mathbf{Y}$ (Metabolomics for example).

And both associated sub-spaces are correlated.

Idea and drawbacks of block normalization

A problem for correlated blocks

If the interesting variance proportions from blocks are different:

- $\mathbf{X}_1$ with $p_1 = 20\ 000$ and only 5% of the variance is associated with $\mathbf{Y}$ (Transciptomics, Spectroscopics,...).
- $\mathbf{X}_2$ with $p_2 = 200$ and 20% of the variance is associated with $\mathbf{Y}$ (Metabolomics for example).

And both associated sub-spaces are correlated.

$$\Longrightarrow \text{ Information from } \mathbf{X}_1 \text{ is hidden by } \mathbf{X}_2, \text{ while } 5\%20\ 000 = 1000 >> 20\%200 = 40.$$

Idea and drawbacks of block normalization

A problem for correlated blocks

If the interesting variance proportions from blocks are different:

- $\mathbf{X}_1$ with $p_1 = 20\ 000$ and only 5% of the variance is associated with $\mathbf{Y}$ (Transciptomics, Spectroscopics,...).
- $\mathbf{X}_2$ with $p_2 = 200$ and 20% of the variance is associated with $\mathbf{Y}$ (Metabolomics for example).

And both associated sub-spaces are correlated.

$$\Longrightarrow \text{ Information from } \mathbf{X}_1 \text{ is hidden by } \mathbf{X}_2, \text{ while}$$
$$5\%20\ 000 = 1000 >> 20\%200 = 40.$$

Additionnal problem in high dimension

For finite n: large p implies over-fitting of models

$$\Longrightarrow \text{ Do regularization.}$$

An unified solution ?

Last observation

Adding useless variables to a block would modify the overall prediction model... a nonsense.

To a new methodology ?

What should it combine ?

- Do not normalize (discard arbitrarily weighting).
- Perform variable selection based on variable marginal correlation with $\mathbf{Y}$: interpretability and regularization.

$$\Longrightarrow \text{ "ddsPLS" .}$$

"ddsPLS" , "PLS" with different covariance matrix estimators

A sparse PLS where sparsity constraints done at covariance estimation step (denoted as $\mathbf{M}^{(r)}$) and not after:

$$\mathbf{S}_\lambda(\mathbf{M}) \quad = \quad \arg\min_{\mathbf{\Sigma}\in\mathbb{R}^{q\times p}} \ \|\mathbf{M}-\mathbf{\Sigma}\|^2 + 2\lambda\,|\mathbf{\Sigma}|, \qquad (1)$$

where λ values are tested along a clever grid and $\mathbf{S}_\lambda$ is the soft-thresholding operator. Interpretation:

- $\lambda = 0$ corresponds to "PLS" model,
- $\lambda = 1$ corresponds to empty model: empirical mean estimation.

"ddsPLS" , the algorithm

The algorithms of "PLS" and "ddsPLS" are close to each other:

"PLS"

$$\begin{cases} \text{(a)} \begin{cases} \mathbf{w}_r = \overrightarrow{\text{RSV}}\left(\mathbf{M}^{(r)}\right), \\ \mathbf{v}_r = \overrightarrow{\text{RSV}}\left(\mathbf{M}^{(r)'}\right), \end{cases} \\ \text{(b)} \quad \mathbf{t}_r = \mathbf{X}^{(r)}\mathbf{w}_r, \\ \text{(c)} \quad \mathbf{p}_r = \mathbf{X}^{(r)'}\mathbf{t}_r/\mathbf{t}_r'\mathbf{t}_r, \\ \text{(d)} \quad \mathbf{c}_r = \mathbf{Y}^{(r)'}\mathbf{t}_r/\mathbf{t}_r'\mathbf{t}_r, \\ \text{(e)} \begin{cases} \mathbf{X}^{(r+1)} = \mathbf{X}^{(r)} - \mathbf{t}_r\mathbf{p}_r', \\ \mathbf{Y}^{(r+1)} = \mathbf{Y}^{(r)} - \mathbf{t}_r\mathbf{c}_r', \end{cases} \end{cases}$$

"ddsPLS"

$$\begin{cases} \text{(a}^\star\text{)} \begin{cases} \mathbf{w}_r = \overrightarrow{\text{RSV}}\left(\mathbf{S}_{\lambda^{(\mathbf{r})}}\left(\mathbf{M}^{(r)}\right)\right), \\ \mathbf{v}_r = \overrightarrow{\text{RSV}}\left(\mathbf{S}_{\lambda^{(\mathbf{r})}}\left(\mathbf{M}^{(r)'}\right)\right), \end{cases} \\ \text{(b)} \quad \mathbf{t}_r = \mathbf{X}^{(r)}\mathbf{w}_r, \\ \text{(c)} \quad \mathbf{p}_r = \mathbf{X}^{(r)'}\mathbf{t}_r/\mathbf{t}_r'\mathbf{t}_r, \\ \text{(d}^\star\text{)} \begin{cases} \mathbf{\Pi}_r = \text{diag}(\delta_{\neq 0}(\mathbf{v}_r)_j) \\ \mathbf{c}_r = \left(\mathbf{Y}^{(r)}\mathbf{\Pi}_r\right)'\mathbf{t}_r/(\mathbf{t}_r'\mathbf{t}_r). \end{cases} \\ \text{(e)} \begin{cases} \mathbf{X}^{(r+1)} = \mathbf{X}^{(r)} - \mathbf{t}_r\mathbf{p}_r', \\ \mathbf{Y}^{(r+1)} = \mathbf{Y}^{(r)} - \mathbf{t}_r\mathbf{c}_r', \end{cases} \end{cases}$$

"ddsPLS" , fix the number of components and the regularization coefficients

Two types of parameters:

- Number R of components,
- regularization parameter per component $(\lambda_r)_{r=1...R}$.

Based on B bootstrap operations for each component:

- Minimize $\bar{R}^2_B - \bar{Q}^2_B$ (minimize over-fitting).
- The $\bar{Q}^2_{B,r}$ must be positive (learn from data).
- The $\bar{Q}^2_B$ is increasing with r (learn something new with the current component).

$\bar{R}^2_B$, $\bar{Q}^2_B$ and $\bar{Q}^2_{B,r}$ are defined in annex but correspond to R^2 and Q^2 at complete model level or component levels in the context of bootstrap considering empirical mean aggregation.

Simulation analysis

Challenge "**ddsPLS**" and "**PLS**" in the high dimensional multiblock context in two cases:

- $\mathbf{X}_t$ blocks are associated with the same sub-space of $\mathbf{Y}$.
- $\mathbf{X}_t$ blocks are associated with different sub-spaces of $\mathbf{Y}$.

Benchmark of 5 approaches

Two blocks are available ($\mathbf{X_1}$ and $\mathbf{X_2}$) with $p_1 >> p_2$. Block $\mathbf{Y}$ such as $q = 1$. Five approaches are compared:

- **(I)**: Predict $\mathbf{Y}$ from $\mathbf{X}_1$ using "**PLS**" ,
- **(II)**: Predict $\mathbf{Y}$ from $\mathbf{X}_2$ using "**PLS**" ,
- **(III)**: Predict $\mathbf{Y}$ from $[\mathbf{X}_1, \mathbf{X}_2]$ using "**PLS**" ,
- **(IV)**: Predict $\mathbf{Y}$ from $[\mathbf{X}_1/\sqrt{p_1}, \mathbf{X}_2/\sqrt{p_2}]$ using "**PLS**" ,
- **(V)**: Predict $\mathbf{Y}$ from $[\mathbf{X}_1, \mathbf{X}_2]$ using "**ddsPLS**" ,

Design 1, Statistical Model

$$
\begin{aligned}
y &= \boldsymbol{\phi_1} + \epsilon \\
\forall j = 1 \dots p_1, \quad x_j^{(1)} &= \begin{cases} \boldsymbol{\phi_1} + \mu_j & \text{if } j = 1 \dots 1\,000, \\ \phi_2 + \mu_j & \text{if } j = 1\,001 \dots 2\,000, \\ \phi_3 + \mu_j & \text{if } j = 2\,002 \dots 3\,000, \\ \mu_j & \text{if } j = 3\,001 \dots 20\,000, \end{cases} \\
\forall j = 1 \dots p_2, \quad x_j^{(2)} &= \begin{cases} \boldsymbol{\phi_1} + \eta_j & \text{if } j = 1 \dots 40, \\ \phi_2 + \eta_j & \text{if } j = 41 \dots 80, \\ \phi_4 + \eta_j & \text{if } j = 81 \dots 120, \\ \eta_j & j = 121 \dots 200, \end{cases}
\end{aligned} \tag{2}
$$

- $\text{var}(y) = \text{var}(x_j^{(1)}) = \text{var}(x_j^{(2)}) = 1$,
- $\text{var}(\phi_1) = \text{var}(\phi_2) = \text{var}(\phi_3) = \text{var}(\phi_4) = \alpha^2 = 0.9$,
- $\text{cov}(\phi_i, \phi_j) = \delta_{i,j}\alpha^2$, $\text{cov}(\phi_i, \epsilon) = \text{cov}(\phi_i, \mu_j) = \text{cov}(\phi_i, \eta_j) = 0$,

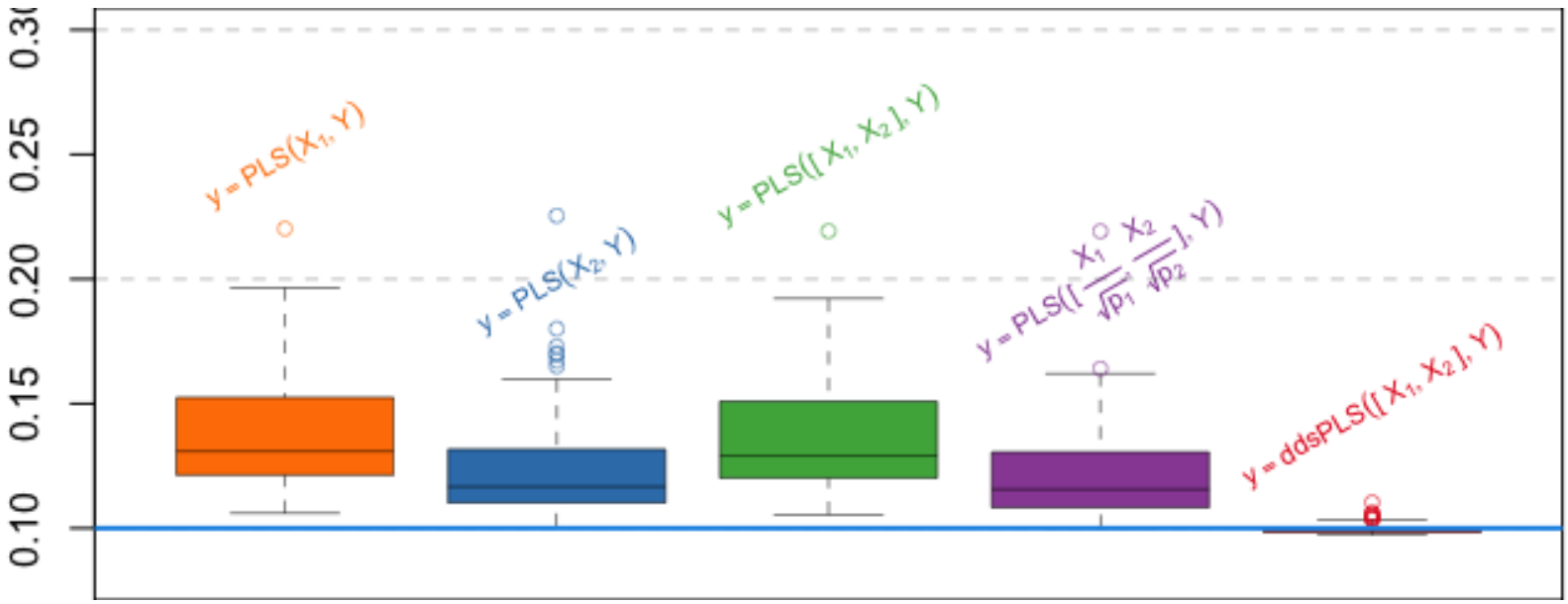
Simulation parameters

- $n = 100$ sampled $N = 100$ times.
- An independent test data-set $n_{test} = 1000$.

Remark

"**ddsPLS**" builds always only 1 component (the objective) and "**PLS**" approaches are constrained to build one component.

MSE on the $n_{test} = 1000$ independent test data set

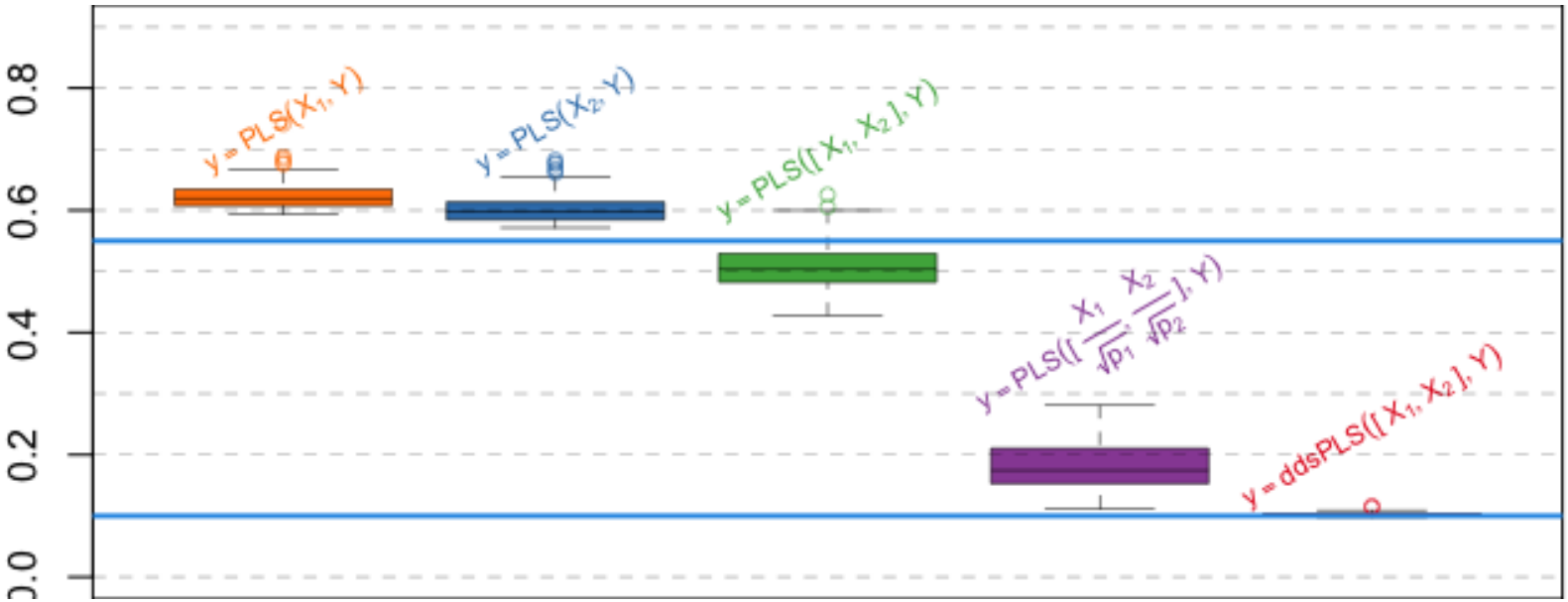


- **(I)** and **(III)** are equivalent: high dimension dominates.
- **(II)** and **(IV)** are equivalent: high SNR dominates.
- "**ddsPLS**" **(V)** deals with high dimension problem with no arbitrary normalization.

Design 2, Statistical Model

$$
\begin{aligned}
y &= (\boldsymbol{\phi_1} + \boldsymbol{\phi_2})/\sqrt{2} + \epsilon \\
\forall j = 1 \dots p_1, \quad x_j^{(1)} &= \begin{cases} \boldsymbol{\phi_1} + \mu_j & \text{if } j = 1 \dots 1\,000, \\ \phi_3 + \mu_j & \text{if } j = 1\,001 \dots 2\,000, \\ \phi_4 + \mu_j & \text{if } j = 2\,002 \dots 3\,000, \\ \mu_j & \text{if } j = 3\,001 \dots 20\,000, \end{cases} \\
\forall j = 1 \dots p_2, \quad x_j^{(2)} &= \begin{cases} \boldsymbol{\phi_2} + \eta_j & \text{if } j = 1 \dots 40, \\ \phi_3 + \eta_j & \text{if } j = 41 \dots 80, \\ \phi_6 + \eta_j & \text{if } j = 81 \dots 120, \\ \eta_j & j = 121 \dots 200, \end{cases}
\end{aligned}
\tag{3}
$$

MSE on the $n_{test} = 1000$ independent test data set



- **(I)** a bit worse than **(II)** due to SNR.
- **(IV)** better than **(III)** hiding $\mathbf{X}_1$ noise due to weighting but cannot properly reconstruct ϕ_1.
- "**ddsPLS**" **(V)** performs better due to regularization and always builds 2 components.

Conclusion over simulations

Performances of traditional normalized multiblock approaches depend on the correlation structure of the data.

It can be wether:

- a relative bad idea (design 1).
- a relative good idea (forgetting high dimensional problem in design 2) but partially hide an important dimension.

In all cases "**ddsPLS**" performs very well in **Prediction**, **Selection** (not shown) and in **Parameter estimation**.

Batch Evolution Modelling

- 5 Chinese Hamster Ovary (CHO) cell cultures performed on univessels.
- Transcriptomics (**Tr**) and Metabolic (**Me**) profiles acquired at 12 time points.

$n = 52$

$p_{\text{Tr}} = 20\ 373$

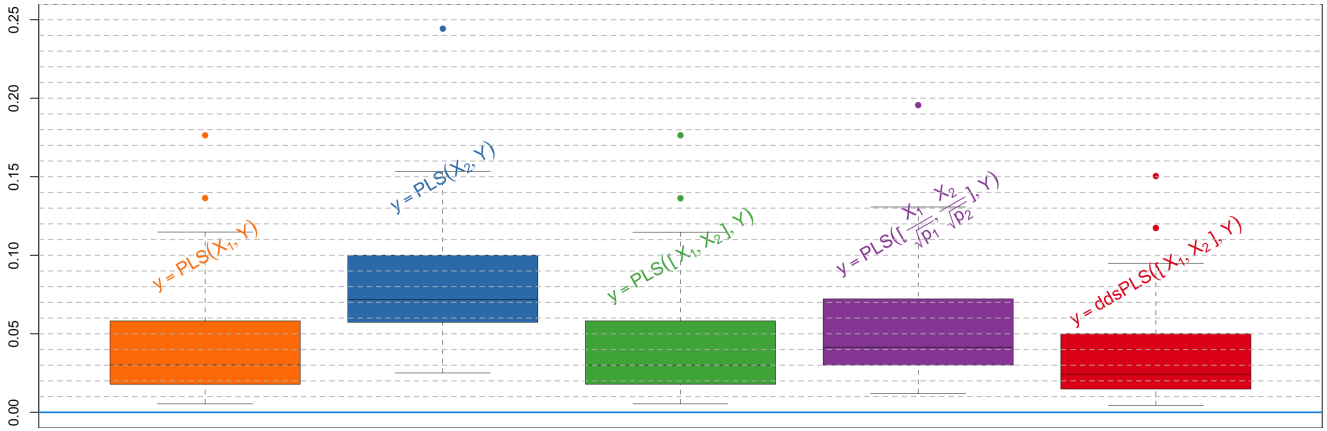
Tr

$p_{\text{Me}} = 58$

Me

Prediction performances (MSE): 50 repetitions (9/10 train, 1/10 test)

Parameter selection (cross-validation): $R_{(I)} = 4$, $R_{(II)} = 2$, $R_{(III)} = 4$, $R_{(IV)} = 4$. And $R_{(V)} = 2$.



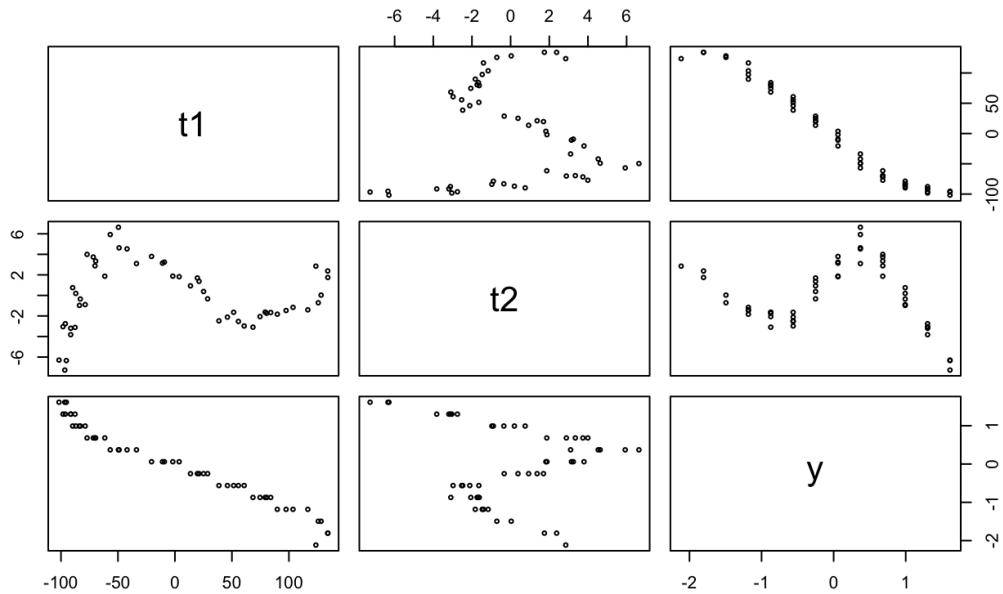
$\mathbf{X}_1$ alone is better than $[\mathbf{X}_1, \mathbf{X}_2]$ or $[\mathbf{X}_1/\sqrt{p_1}, \mathbf{X}_2/\sqrt{p_2}]$: $\mathbf{X}_1$ builds many interesting components through a good SNR.

"**ddsPLS**" **(V)** performs slightly better: regularization effect.

"ddsPLS" performances

Variable selection: **Tr** 48% and **Me** 62%. Two components built:

- Dim 1: 97% variance explained.
- Dim 2: 2% variance explained.



Scores versus y... Non linearities ?

Conclusion

When to use block normalization ?

- Low and equivalent dimensions from one block to another.
- Independent blocks.
- Equivalent "signal to noise ratio" from blocks to blocks.
- . . . so, why using it ?

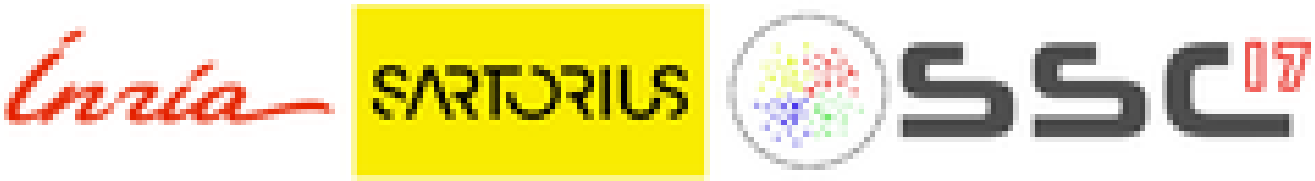
When to **NOT** use block normalization ?

- High dimension.
- Different dimensions in different orders of magnitude.
- Dependent blocks.

"**ddsPLS**" (github.com/hlorenzo/ddsPLS2) selects variables based on their marginal correlations: no need to normalize blocks.

Thanks

INRIA & Sartorius & SSC17 Conference Committee



Quality criteria

$$\bar{R}^2_B = \frac{1}{B}\sum_{b=1}^{B} R^2_b \quad \text{and} \quad \bar{Q}^2_B = \frac{1}{B}\sum_{b=1}^{B} Q^2_b \tag{4}$$

with, for the current bootstrap sample b,

$$\begin{aligned} R^2_b &= 1 - \frac{\sum_{j=1}^{q}\sum_{i\in \mathsf{IN}(b)}\left(y_{i,j} - \hat{y}^b_{i,j}\right)^2}{\sum_{j=1}^{q}\sum_{i\in \mathsf{IN}(b)}\left(y_{i,j} - \bar{y}^b_{j}\right)^2}, \\ Q^2_b &= 1 - \frac{\sum_{j=1}^{q}\sum_{i\in \mathsf{OOB}(b)}\left(y_{i,j} - \hat{y}^b_{i,j}\right)^2}{\sum_{j=1}^{q}\sum_{i\in \mathsf{OOB}(b)}\left(y_{i,j} - \bar{y}^b_{j}\right)^2}, \end{aligned} \tag{5}$$

Quality criteria (2)

In the same way, bootstrapped versions of R^2_r and Q^2_r are given by

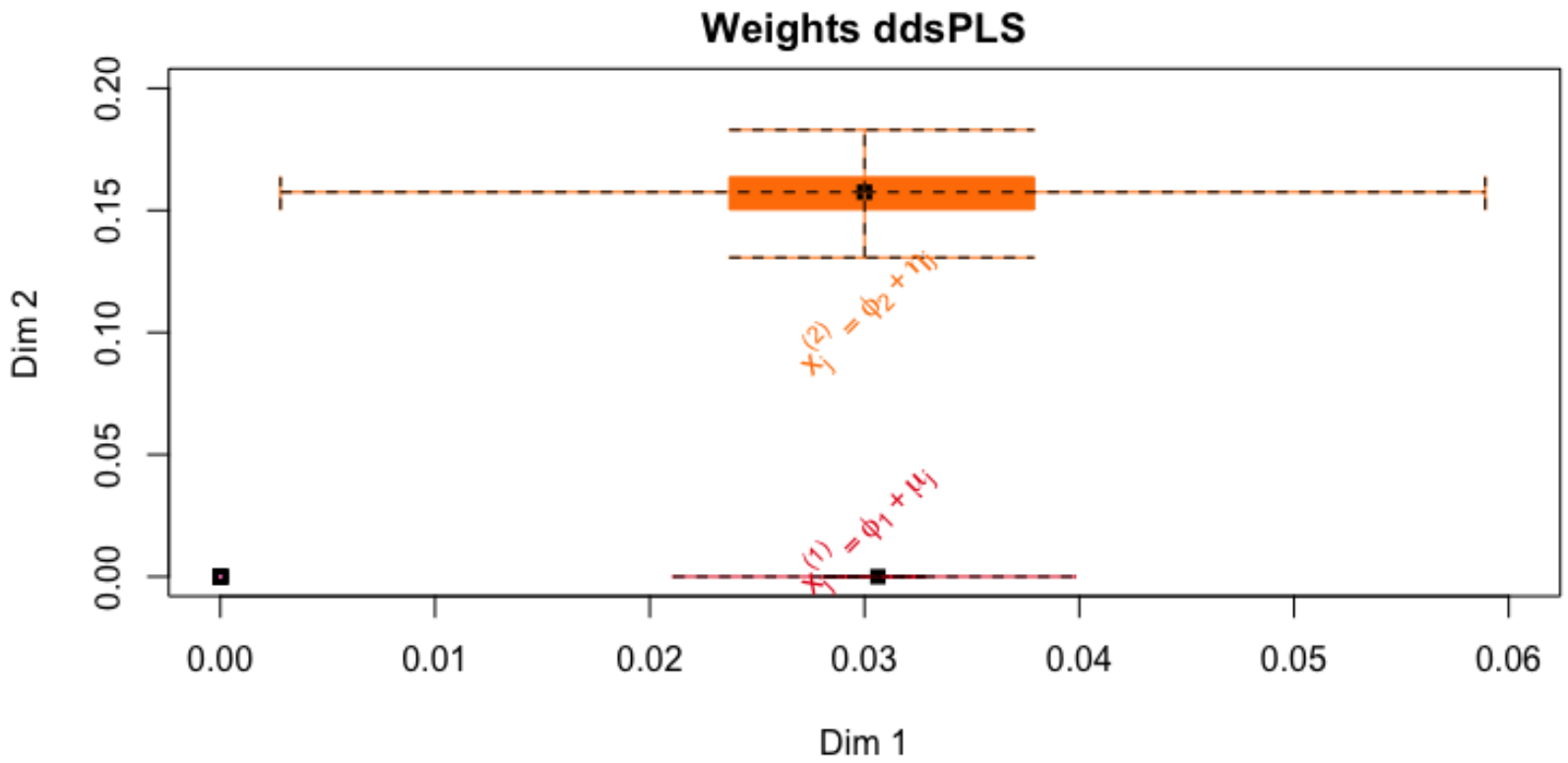
$$\bar{R}^2_{B,r} = \frac{1}{B}\sum_{b=1}^{B} R^2_{b,r} \quad \text{and} \quad \bar{Q}^2_{B,r} = \frac{1}{B}\sum_{b=1}^{B} Q^2_{b,r} \tag{6}$$

where

$$\begin{aligned} R^2_{b,r} &= 1 - \frac{\sum_{j=1}^{q}\sum_{i\in \mathsf{IN}(b)}\left(y_{i,j} - \left(\hat{y}^{b,(r)}_{i,j} - \hat{y}^{b,(r-1)}_{i,j}\right) - \bar{y}^b_j\right)^2}{\sum_{j=1}^{q}\sum_{i\in \mathsf{IN}(b)}\left(y_{i,j} - \bar{y}^b_j\right)^2}, \\ Q^2_{b,r} &= 1 - \frac{\sum_{j=1}^{q}\sum_{i\in \mathsf{OOB}(b)}\left(y_{i,j} - \hat{y}^{b,(r)}_{i,j}\right)^2}{\sum_{j=1}^{q}\sum_{i\in \mathsf{OOB}(b)}\left(y_{i,j} - \hat{y}^{b,(r-1)}_{i,j}\right)^2}. \end{aligned} \tag{7}$$

Variable Selection design 1, "ddsPLS"

Variable Selection design 2, "ddsPLS"



References I

Westerhuis, J., Kourti, T., & MacGregor, J. (1998) Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, **12**.

Wold, H. (1966) Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391–420.

Wold, S., Kettaneh, N., & Tjessem, K. (1996) Hierarchical multiblock pls and pc models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, **10**, 463–482.