ANALYSE SUPERVISÉE MULTIBLOC EN GRANDE DIMENSION

HADRIEN LORENZO *Postdoc in SISTM team,*

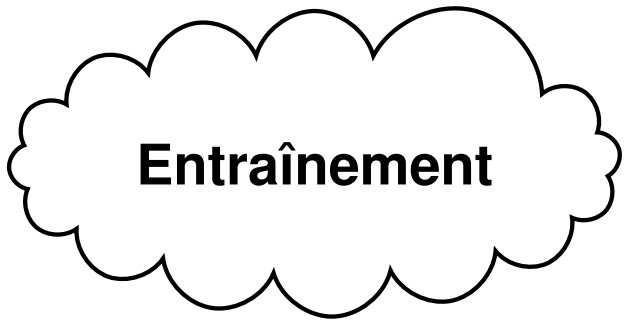
INSERM/INRIA Bordeaux

MARDI 21 JANVIER 2020

Contents

- Une introduction à l'apprentissage statistique
  - Décomposition du risque
  - Compromis biais-variance
  - Le problème linéaire
  - Solutions envisagées
  - La régularisation
- My thesis
- ddsPLS, complete data
  - Statistical model and estimators
  - Application to a real data set
- Koh-Lanta, missing values per block
  - Algorithm
  - Simulations
  - Application to real data sets
    - Ebola data set
    - Venous thrombosis data set
- Packages

Modéliser l'apprentissage statistique



- $\mathcal{D}_n$ un jeu de données d'entraînement formé de $n > 0$ réalisations indépendantes du couple (X, Y),
- X est de dimension $p > 0$ et Y est de dimension $q > 0$,
- les couples $(X_i, Y_i)_{i \in [\![1,n]\!]}$ sont supposées suivre une même loi $\mathbb{P}$ inconnue.

Modèles

On suppose un modèle de lien statistique entre X et Y de la forme

$$Y = f(X) + \epsilon,$$

où ϵ est le bruit d'observation, centré et de variance σ^2_ϵ et f est supposée déterministe.

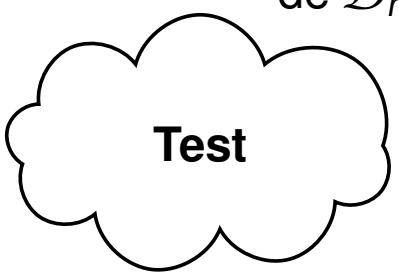
On produit $M > 0$ estimateurs de la fonction f sur les données $\mathcal{D}_n$ que l'on note $\left(\hat{f}_n^{(m)}\right)_{m \in [\![1,M]\!]}$, dans des espaces fonctionnels définis par l'analyste et les experts.

Estimateurs

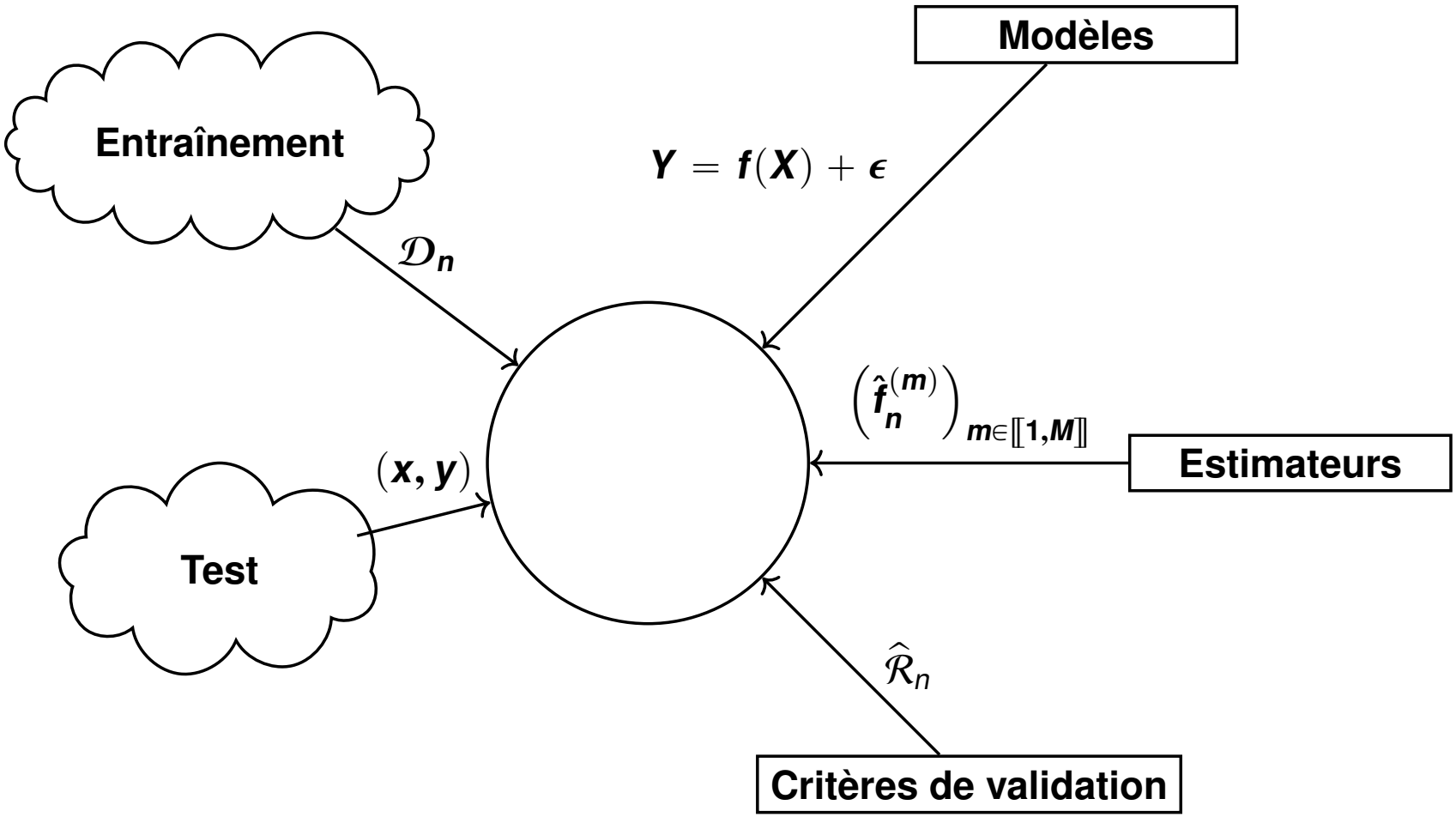
On compare les modèles grâce à un risque empirique $\widehat{\mathcal{R}}_n$ tel que

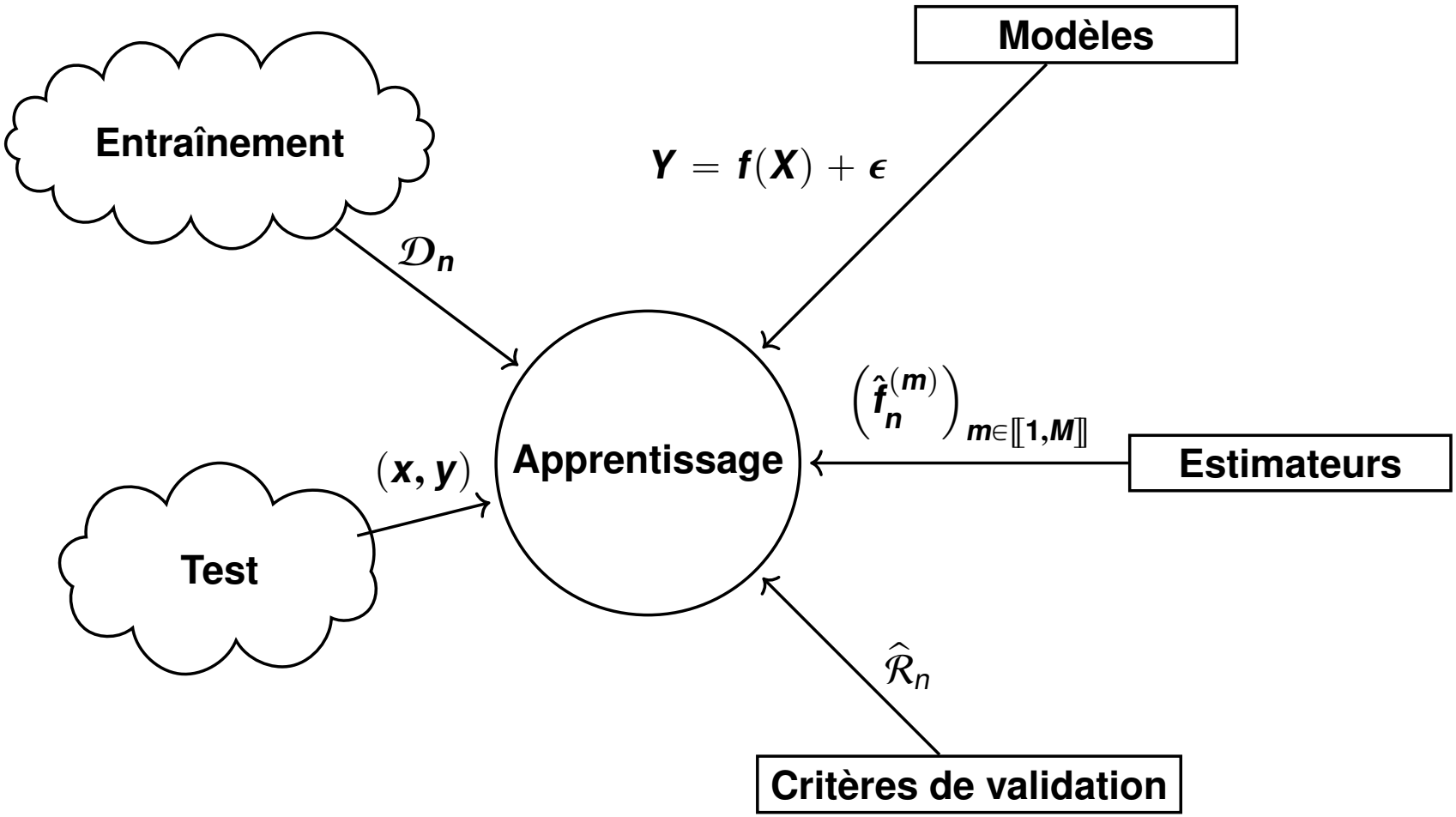
$$\hat{f}_{opt} = \underset{m \in [\![1,M]\!]}{\arg\max}\, \widehat{\mathcal{R}}_n \left(\hat{f}_n^{(m)} \right).$$

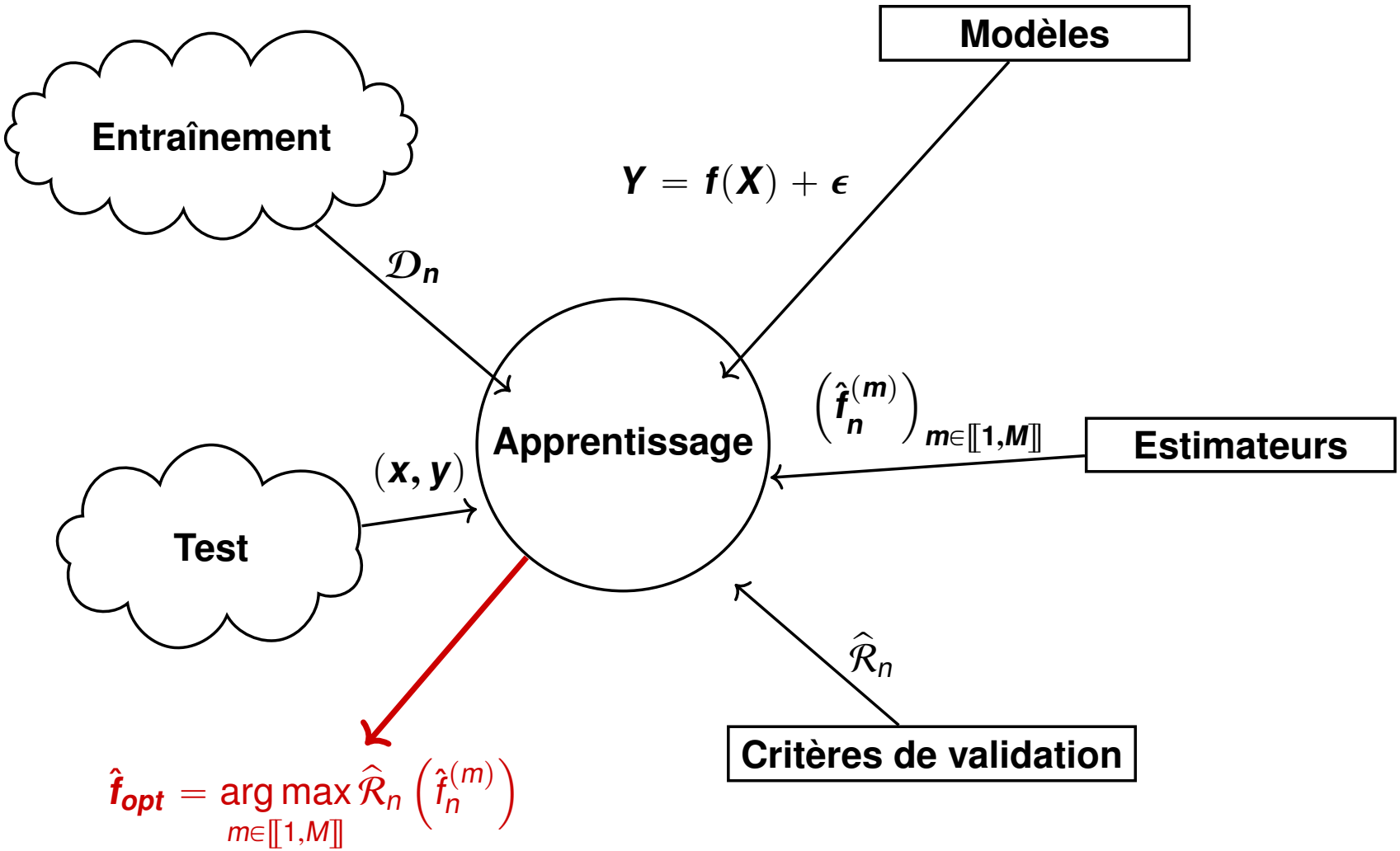
Ce risque est souvent calculé sur un échantillon (x, y) indépendant de $\mathcal{D}_n$ et appelé échantillon de test.



Critères de validation


Modèles
Entraînement
$Y = f(X) + \epsilon$
$\mathcal{D}_n$
$\left(\hat{f}_n^{(m)}\right)_{m \in [\![1,M]\!]}$
Estimateurs
(x, y)
Test
$\widehat{\mathcal{R}}_n$
Critères de validation


Modèles
Entraînement
$Y = f(X) + \epsilon$
$\mathcal{D}_n$
$\left(\hat{f}_n^{(m)}\right)_{m \in [\![1,M]\!]}$
Apprentissage
Estimateurs
(x, y)
Test
$\widehat{\mathcal{R}}_n$
Critères de validation


Modèles
Entraînement
$Y = f(X) + \epsilon$
$\mathcal{D}_n$
$\left(\hat{f}_n^{(m)}\right)_{m\in[\![1,M]\!]}$
Estimateurs
Apprentissage
(x, y)
Test
$\widehat{\mathcal{R}}_n$
Critères de validation
$\hat{f}_{opt} = \underset{m\in[\![1,M]\!]}{\arg\max}\, \widehat{\mathcal{R}}_n\left(\hat{f}_n^{(m)}\right)$

Risque et compromis

Pour toute fonction $f \in \mathcal{F}$, [Vapnik 1992] a décrit ce que l'on appelle le **risque fonctionnel** selon

$$\mathcal{R}(f) = \mathbb{E}\left(L\left(y, f(x)\right)\right) = \int_{\mathcal{X}\times\mathcal{Y}} L\left(y, f(x)\right) Pr(x, y) dxdy, \qquad (1)$$

avec par exemple la *perte quadratique* $L : (y, \hat{y}) \rightarrow ||y - \hat{y}||_2^2$ en régression ou la *hinge loss*, $L : (y, \hat{y}) \rightarrow max(0, 1 - y\hat{y})$, en classification.
Pour des échantillons (car $\mathcal{D}_n$ de taille finie) on définit le **risque empirique**

$$\hat{\mathcal{R}}_n(f) = \mathbb{E}_n\left[L\left(y, f(x)\right) | \{x, y\} \in \mathcal{D}_n\right] = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, f(x_i)\right), \qquad (2)$$

Principe de minimisation du risque empirique (principe ERM) [Vapnik et Chervonenkis 2015]

$\forall \mathcal{D}_n$, la fonction $\hat{f}_n$ qui minimise le risque empirique, est une bonne approximation de f, fonction qui minimise l'équation (1).
Vapnik propose la majoration (qui dépend, de façon croissante de la complexité de l'espace des fonctions et de façon décroissante du nombre d'individus au travers de la fonction ϕ) ce qui est formulé grâce à l'équation

$$\mathcal{R}(f) < \hat{\mathcal{R}}_n(f) + \phi\left(\sqrt{\frac{h}{n}}, \eta\right), \tag{3}$$

où h se rapporte à la complexité de la fonction f et η la probabilité associée à ce risque empirique.

Décomposition du risque

On désigne par *oracle* la fonction $f^\star$ qui est solution du problème dans l'hypothèse où Pr est connue et $\forall L$, on définit

$$\text{la fonction } \textbf{oracle} \text{ par } \quad f^\star \quad = \quad \arg\min_{f \in \mathcal{F}^\star} \mathcal{R}(f), \tag{4}$$

où l'ensemble fonctionnel $\mathcal{F}^\star$ est inconnu mais comprend la fonction $f^\star$ avec de plus $\mathcal{F} \subset \mathcal{F}^\star$.

Décomposition du risque

L'oracle est, bien entendu, inconnue puisque l'échantillon $\mathcal{D}_n$ est fini et que $\mathcal{F}^\star$ est inconnu. Nous introduisons

$$\text{la fonction } \textbf{oracle sur } \mathcal{F} \text{ par } g^\star = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f). \tag{4}$$

Décomposition du risque

L'oracle est, bien entendu, inconnue puisque l'échantillon $\mathcal{D}_n$ est fini et que $\mathcal{F}^\star$ est inconnu. Nous introduisons

$$\text{la fonction } \textbf{oracle sur } \mathcal{F} \text{ par } g^\star = \arg\min_{f \in \mathcal{F}} \mathcal{R}(f). \tag{4}$$

On peut alors décomposer

$$\mathcal{R}\left(\hat{f}_n\right) - \mathcal{R}\left(f^\star\right) = \mathcal{R}\left(\hat{f}_n\right) - \mathcal{R}\left(g^\star\right) + \mathcal{R}\left(g^\star\right) - \mathcal{R}\left(f^\star\right) .$$

La première différence à droite de l'égalité décrit l'**erreur d'estimation** et la seconde décrit l'**erreur d'approximation**.

Compromis biais-variance

Soit $L : ||.||_2^2$ la perte quadratique, [Geman, Bienenstock et Doursat 1992] ont démontré que l'espérance du risque de $\hat{f}_n$ peut se décomposer en trois termes

$$\mathbb{E}\left[\mathcal{R}\left(\hat{f}_n\right)\right] = \text{bruit}\left(\hat{f}_n\right) + \text{biais}^2\left(\hat{f}_n\right) + \text{var}_{est}\left(\hat{f}_n\right), \quad (5)$$

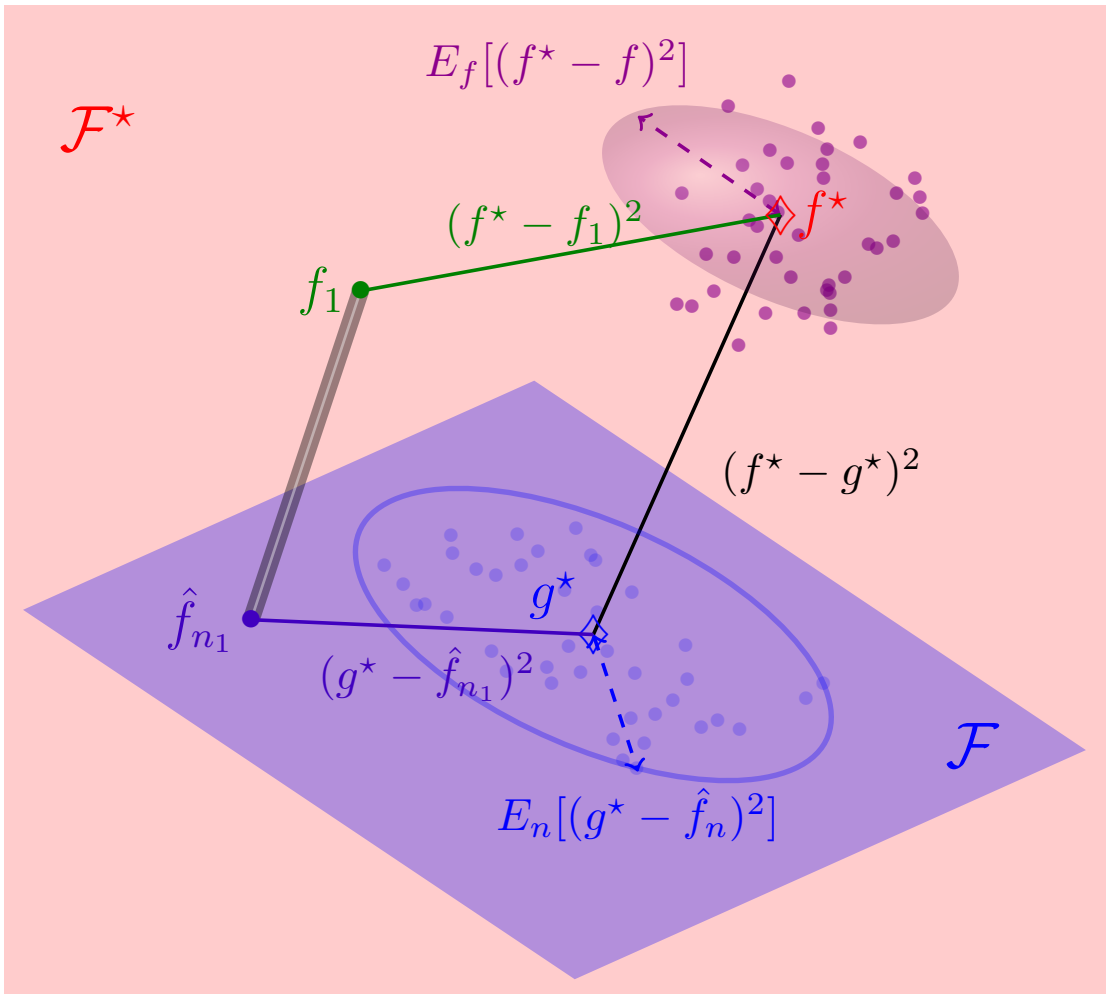
où le terme bruit $\left(\hat{f}_n\right)$ est irréductible, c'est l'erreur d'observation.

Compromis biais-variance

Termes d'erreur du risque empirique

$$\begin{aligned}
\text{Erreur de bruit} \quad &: \quad \widehat{\text{bruit}}\left(\hat{f}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(f\left(x_i\right) - f^{\star}\left(x_i\right)\right)^2 \\
\text{Erreur d'approximation} \quad &: \quad \widehat{\text{biais}^2}\left(\hat{f}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(f^{\star}(x_i) - g^{\star}(x_i)\right)^2 \\
\text{Erreur d'estimation} \quad &: \quad \widehat{\text{var}}\left(\hat{f}_n\right) = \frac{1}{n}\sum_{i=1}^{n}\left(g^{\star}(x_i) - \hat{f}_n(x_i)\right)^2
\end{aligned} \tag{5}$$

Compromis biais-variance



Cas particulier du modèle linéaire

On suppose qu'il existe $\mathbf{b} \in \mathbb{R}^p$ associé au vecteur aléatoire y.
$\mathbf{b}$ est supposé déterministe et on suppose une erreur de mesure ϵ centrée de variance σ^2_ϵ additive.
On écrit le modèle de mesure suivant

$$y = X^T\mathbf{b} + \epsilon,$$

Soit en notation matricielle

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}. \tag{5}$$

Cas particulier du modèle linéaire

$\mathbf{x}_1$, une observation de test, $f^\star(\mathbf{x}_1)$ et $f(\mathbf{x}_1)$ sont :

$$\begin{aligned} f^\star(\mathbf{x}_1) &= \mathbf{x}_1\mathbf{b}, \\ f(\mathbf{x}_1) &= \mathbf{x}_1\mathbf{b} + \epsilon_1, \end{aligned}$$

en notant ϵ_1 le bruit d'observation associé.

Cas particulier du modèle linéaire

Le problème des **moindres carrés ordinaires** minimise la perte quadratique du modèle (5), donc

$$\hat{\mathbf{b}}_n^{(MCO)} = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 .$$

Ce problème a pour solution

$$\hat{\mathbf{b}}_n^{(MCO)} = \mathbf{X}^+\mathbf{y}, \tag{5}$$

où l'on note $\mathbf{X}^+$ la pseudo-inverse de Moore-Penrose de $\mathbf{X}$.
Sur un échantillon de test $\mathbf{x}_1$, $g^\star(\mathbf{x}_1)$ et $\hat{f}_n^{(MCO)}(\mathbf{x}_1)$ s'écrivent :

$$\begin{aligned} g^\star(\mathbf{x}_1) &= \mathbf{x}_1\mathbf{b}, \\ \hat{f}_n^{(MCO)}(\mathbf{x}_1) &= \mathbf{x}_1\mathbf{b} + \mathbf{x}_1\mathbf{X}^+\boldsymbol{\epsilon} \end{aligned}$$

Cas particulier du modèle linéaire

$$\begin{aligned}
\text{Erreur de bruit} &: \text{bruit}\left(\hat{b}_n^{(MCO)}\right) = \sigma_\epsilon^2 \\
\text{Erreur d'approximation} &: \text{biais}^2\left(\hat{b}_n^{(MCO)}\right) = 0 \\
\text{Erreur d'estimation} &: \text{var}_{est}\left(\hat{b}_n^{(MCO)}\right) = p\sigma_\epsilon^2 \text{Trace}\left(\mathbf{X}^{+T}\mathbf{X}^{+}\right)
\end{aligned} \tag{5}$$

L'espérance du risque empirique s'écrit comme la somme des trois termes associés aux trois erreurs, soit

$$\mathbb{E}\left[\mathcal{R}\left(\hat{b}_n^{(MCO)}\right)\right] = \sigma_\epsilon^2 + p\sigma_\epsilon^2 \text{Trace}\left(\mathbf{X}^{+T}\mathbf{X}^{+}\right).$$

Cas particulier du modèle linéaire

Donc :

- l'estimateur des moindres carrés est sans biais,
- la variance de l'estimateur des moindres carrés diverge si la matrice échantillon $\mathbf{X}$ n'est plus inversible.

Le **conditionnement** (rapport de la plus grande valeur singulière sur la plus petite par exemple) permet d'évaluer la difficulté d'inverser numériquement une matrice.
Si le conditionnement de la matrice diverge, alors la variance de l'estimateur des moindres carrés diverge aussi.

Cas particulier du modèle linéaire

En effet :
Si l'on suppose un design orthogonal qui correspond à $\mathbf{X}^T\mathbf{X} = n\mathbb{I}_p$,
il vient que Trace $\left(\mathbf{X}^{+T}\mathbf{X}^{+}\right) = 1$ et alors

$$\mathbb{E}\left[\mathcal{R}_{ortho}\left(\hat{\mathbf{b}}_n^{(MCO)}\right)\right] = \sigma_\epsilon^2(1+p). \quad (5)$$

Le risque quadratique est donc une fonction affine de la complexité du modèle linéaire représentée par p.

Grande dimension

La **grande dimension** regroupe tous les cas de mauvais conditionnement de **X**.

Solution « Brut Force »

Tester tous les modèles.
Si le modèle nécessite l'estimation de $k \leqslant n$ paramètres pour p variables.

$$\longrightarrow C_p^k \text{ possibilités.}$$

Limitations pour tester tous les modèles.

Solutions pas à pas

Retirer ou ajouter une variable utile pour le modèle courant.
Nécessité de créer un critère de coût définissant l'utilité d'une variable (voir les *forêts aléatroires*).

La régularisation

Principe

Pénaliser l'optimisation des paramètres en contraignant leur amplitude à ne pas être trop importante.
Le risque empirique pénalisé s'écrit

$$\min_{f \in \mathcal{F}} \hat{\mathcal{R}}_n(f) + \lambda \|f\|_{\mathcal{F}}, \tag{6}$$

Où $\|f\|_{\mathcal{F}}$ mesure l'amplitude de la focntion f.

La régularisation

Idée et lien avec la décomposition biais-variance

- Ajout de quelques paramètres (λ,..) qui ne sont pas dans le modèle, donc augmentation de l'erreur d'approximation, le biais.
- Réduction de l'erreur d'estimation, car l'estimation oublie des caractéristiques individuelles et se généralise mieux.

La régularisation Ridge

Problème, solution et oracles

Soit le problème régularisé

$$\min_{\mathbf{b}\in\mathbb{R}^p} \hat{\mathcal{R}}_n(\mathbf{b}) + \lambda\|\mathbf{b}\|_2^2, \tag{6}$$

qui a pour solution, $\forall\lambda > 0$,

$$\hat{\mathbf{b}}_n^{(Ridge)} = \left(\mathbf{X}^T\mathbf{X}/n + \lambda\mathbb{I}\right)^{-1}\mathbf{X}^T\mathbf{y}/n = \mathbf{X}^{(+,\lambda)}\mathbf{y},$$

où $\mathbf{X}^{(+,\lambda)} = \left(\mathbf{X}^T\mathbf{X}/n + \lambda\mathbb{I}\right)^{-1}\mathbf{X}^T/n$. On peut donc écrire

$$f(\mathbf{x}_1) = y_1 = \mathbf{x}_1\mathbf{b} + \epsilon_1, \qquad f^\star(\mathbf{x}_1) = \mathbf{x}_1\mathbf{b},$$
$$g^\star(\mathbf{x}_1) = \mathbb{E}_n\left[\mathbf{x}_1\hat{\mathbf{b}}_n^{(Ridge)}\right], \qquad \hat{f}_n(\mathbf{x}_1) = \mathbf{x}_1\hat{\mathbf{b}}_n^{(Ridge)},$$

La régularisation Ridge

Termes d'erreur

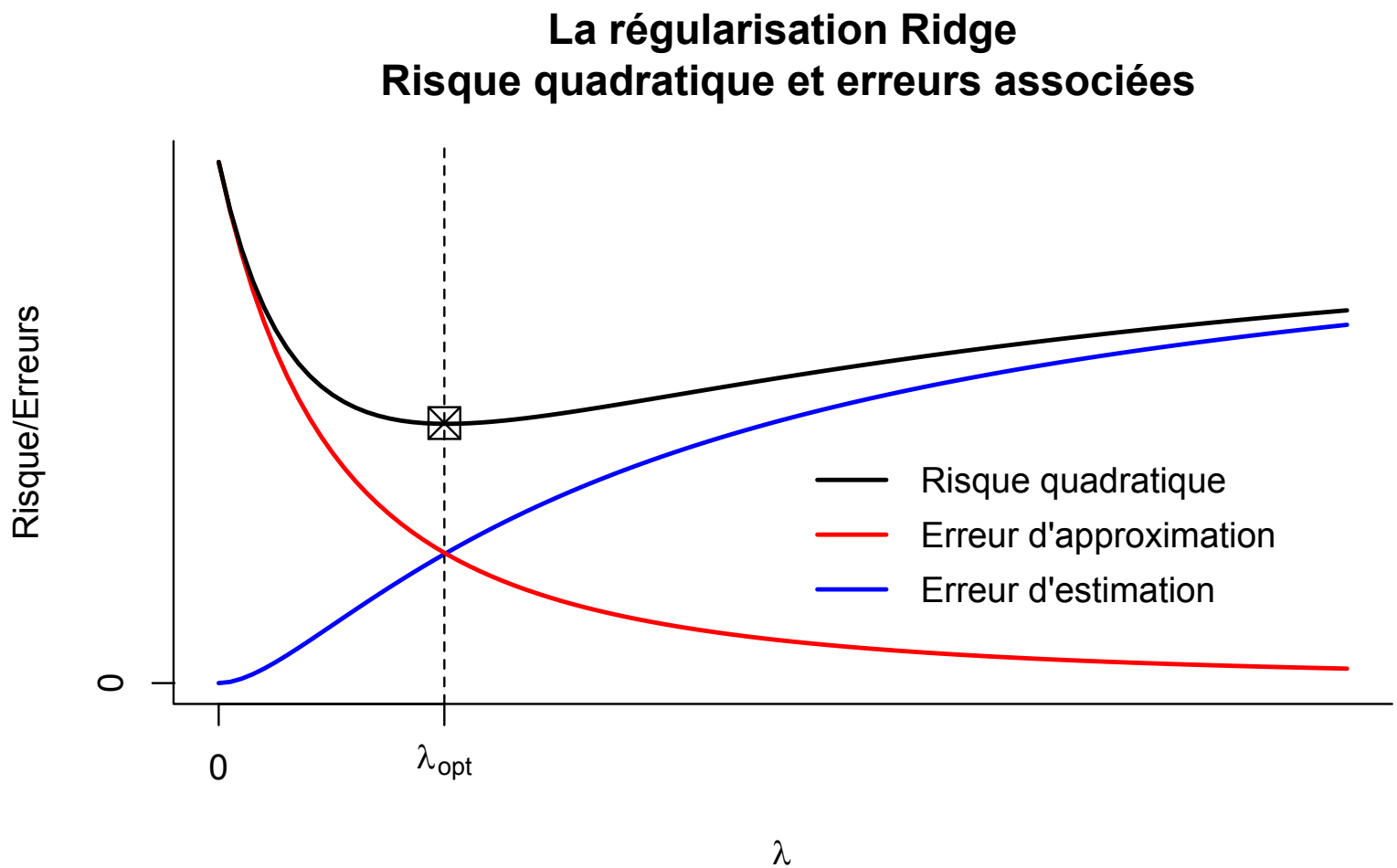
Les termes d'erreur du risque quadratique pour cet estimateur sont

$$\begin{aligned}
\text{Erreur de bruit} &: \text{bruit}\left(\hat{b}_n^{(Ridge)}\right) = \sigma_\epsilon^2 \\
\text{Erreur d'approximation} &: \text{biais}^2\left(\hat{b}_n^{(Ridge)}\right) = \lambda^2 \mathbf{b}^T \left(\mathbf{X}^T\mathbf{X}/n + \lambda\mathbb{I}\right)^{-2} \mathbf{b} \\
\text{Erreur d'estimation} &: \text{var}_{est}\left(\hat{b}_n^{(Ridge)}\right) = {\sigma_\epsilon}^2 \text{Trace}\left(\left(\mathbf{X}^T\mathbf{X}/n + \lambda\mathbb{I}\right)\right.
\end{aligned} \tag{6}$$

Et donc

$$\hat{\mathcal{R}}\left(\hat{\mathbf{b}}_n^{(Ridge)}\right)_{orthogonal} = {\sigma_\epsilon}^2 \quad + \quad \mathbf{b}^T\mathbf{b}\frac{\lambda^2}{(1+\lambda)^2} \quad + \quad {\sigma_\epsilon}^2 p \frac{1}{(1+\lambda)^2}. \tag{7}$$

La régularisation Ridge



La régularisation Lasso

Problème et solutions ?

Soit le problème régularisé, $\forall \lambda > 0$,

$$\min_{\mathbf{b} \in \mathbb{R}^p} \hat{\mathcal{R}}_n(\mathbf{b}) + \lambda \|\mathbf{b}\|_1. \tag{6}$$

Solution dans le cas du design orthogonal :

$$\hat{b}_j^{(Lasso)} = \begin{cases} \hat{b}_j^{(MCO)} - \lambda \text{sign}\left(\hat{b}_j^{(MCO)}\right) & si \left|\hat{b}_j^{(MCO)}\right| > \lambda \\ 0 & sinon \end{cases}, \tag{7}$$

L'estimateur réduit les coefficients des moindres carrés ordinaires de façon linéaire d'un coefficient λ jusqu'à tous les annuler.

Solution dans le cas général :

Pas de solution analytique mais algorithmiques (LARS par exemple, méthode pas à pas)

Comparaison Ridge et Lasso

Objectifs et intérêts

- **Ridge** : Réduire quadratiquement le spectre de la matrice de covariance.
- **Lasso** : Réduire linéairement la taille de chaque coefficient.

Le Lasso permet une **sélection de variables** que ne permet pas le Ridge.

Comparaison Ridge et Lasso 2

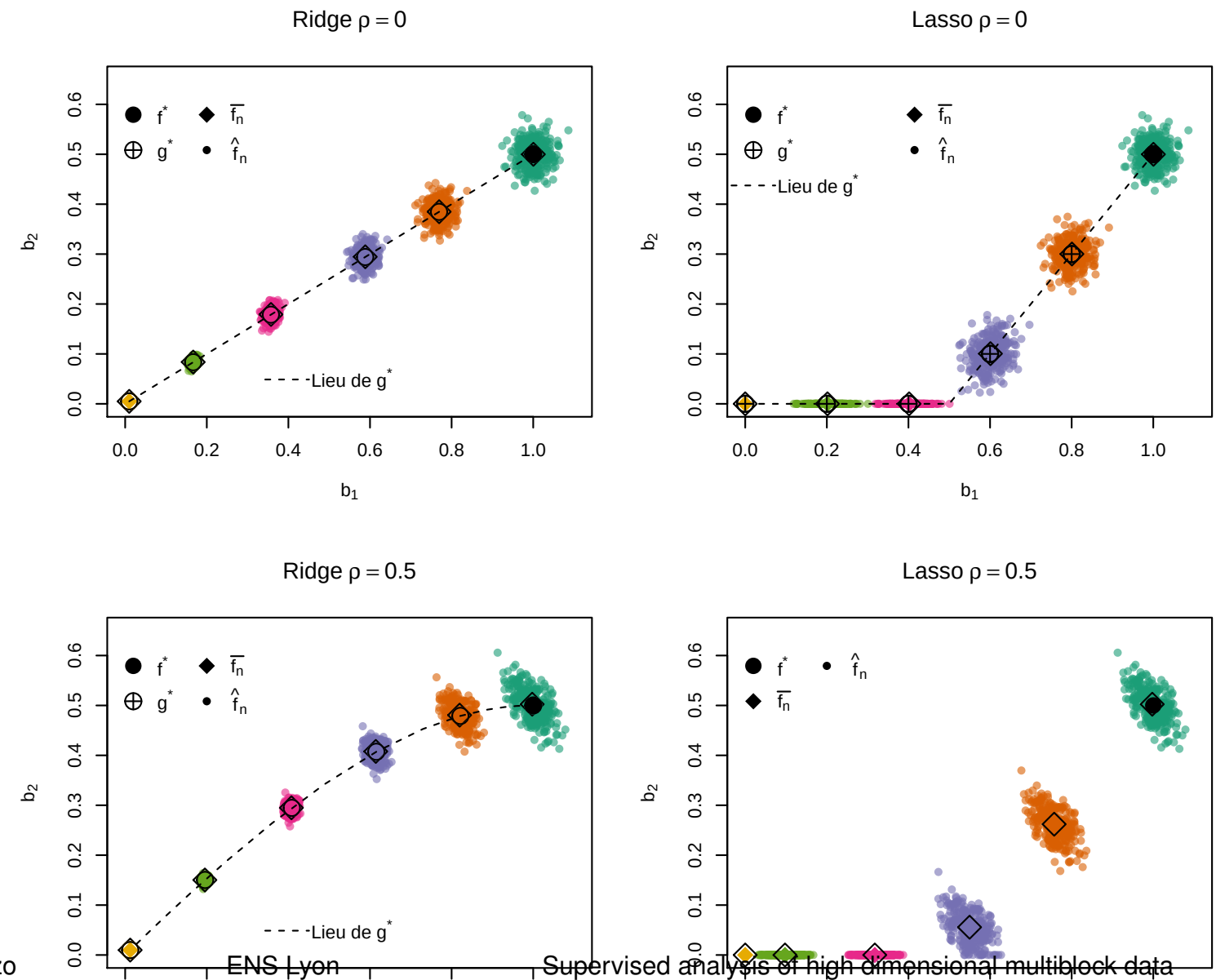
Exercice de simulation

Soit le modèle

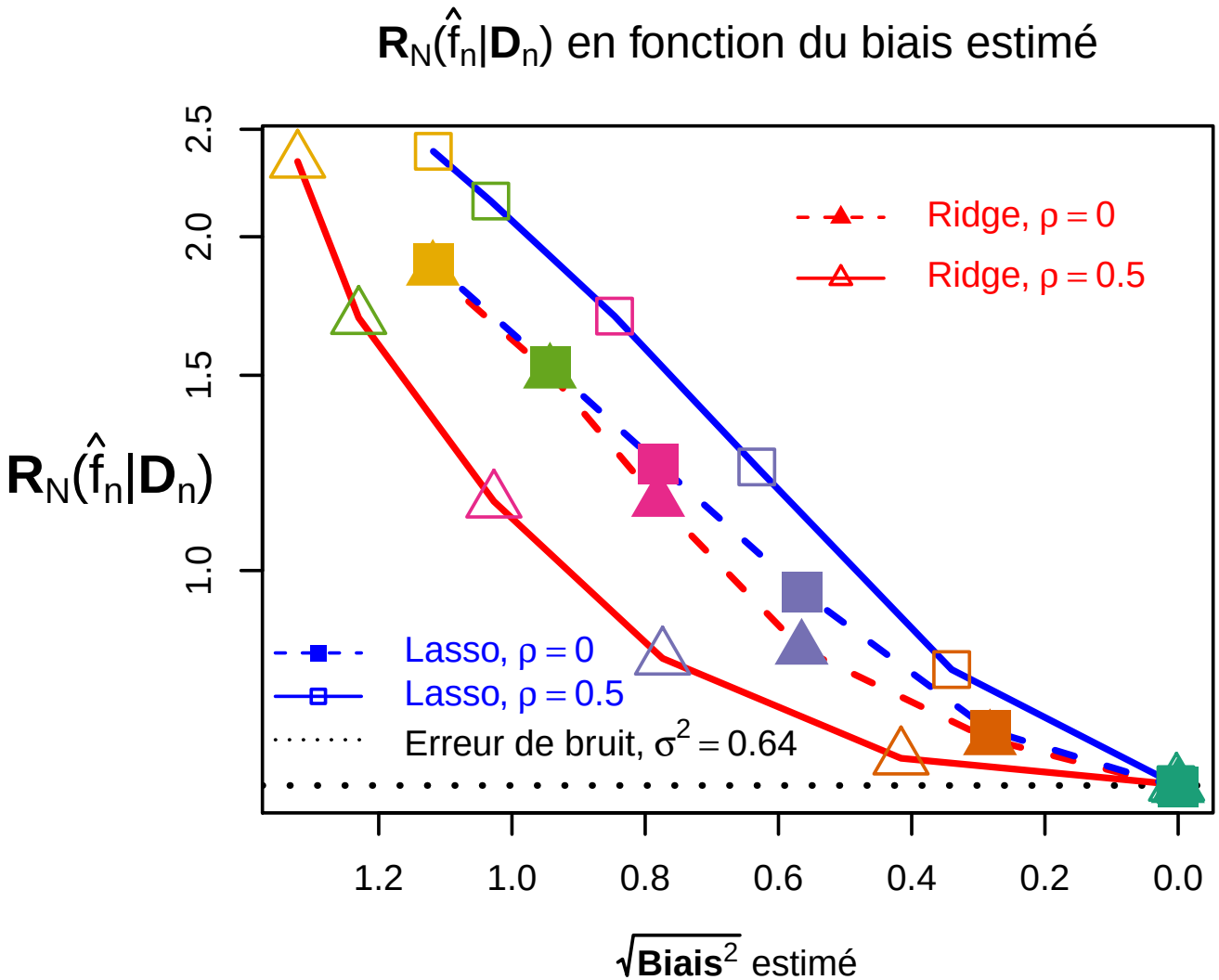
$$Y = b_1 X_1 + b_2 X_2 + \epsilon,$$

où $\epsilon \sim \mathcal{N}(0, 0.8^2)$, $b^\star = (1, 0.5)^T$ et $n = 1000$.
On tire aléatoirement $N = 300$ échantillons pour lesquels on teste différentes paramètres de régularisation.
On suppose de plus que $X_1 \sim \mathcal{N}(0, 1)$ et $X_2 \sim \mathcal{N}(0, 1)$ et une corrélation théorique $\rho = 0$ puis $\rho = 0.5$.

Comparaison Ridge et Lasso 2



Comparaison Ridge et Lasso 2



La méthode PLS [Wold father & son, 1983]

Equivalent à une recherche de sous éléments singuliers, ou *Singular Value Decomposition* (**SVD**), avec déflation. Avec les notations :

- **Poids** u et v : intérêt donné à chaque variable pour X, via u, et pour Y, via v.
- **Scores** ou **variates** des **composantes (principales)** Xu etand Yv : projections de X et Y dans les sous-espaces definis par u et v.

$\Longrightarrow$ Recherches, par projections, dans X de l'information associée à Y.

Résolution du problème PLS

En formalisme lagrangien :

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

$\mathbf{X}_{n \times p}$ et $\mathbf{Y}_{n \times q}$ les matrices, centrées, des covariables et des réponses, alors :

System $\partial_\cdot = 0$:

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

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

Déflation :

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

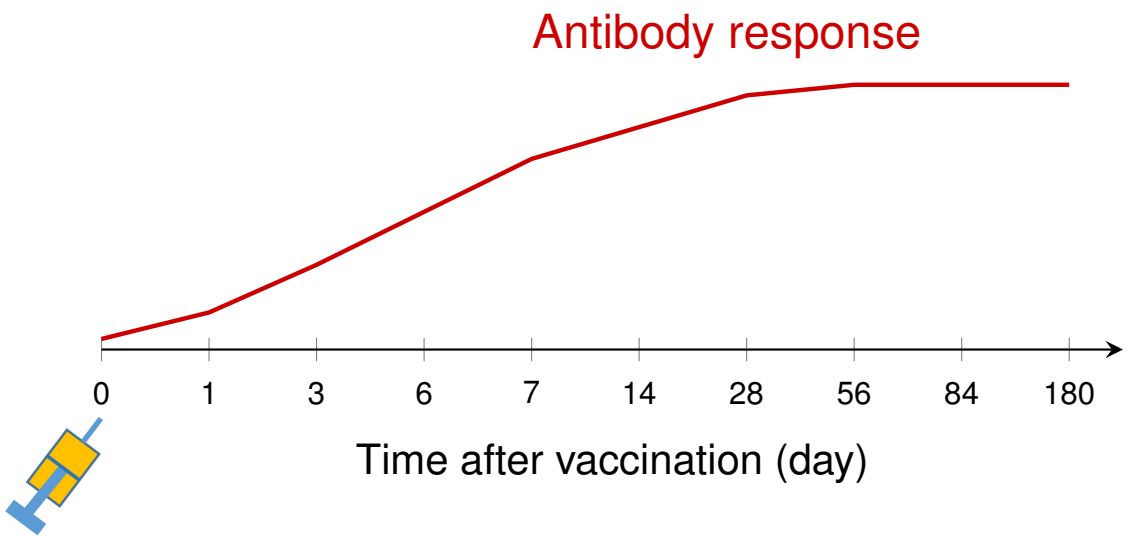
Régression :

$$\mathbf{Y} \approx \mathbf{X}\mathbf{B}$$

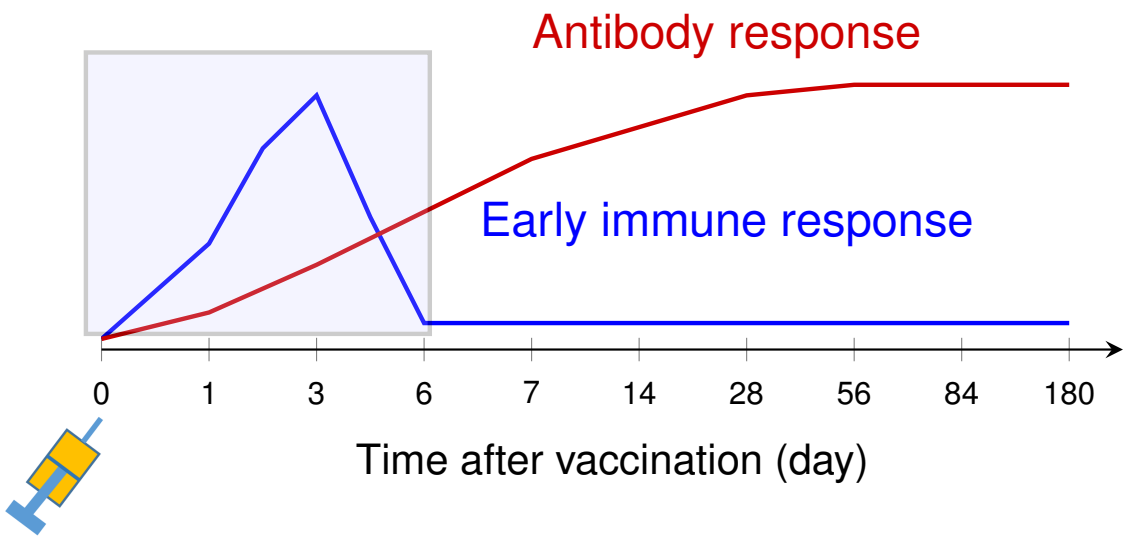
$$\mathbf{B} = \frac{v^T \mathbf{Y}^T \mathbf{X} u}{||\mathbf{X} u||_2^2} u v^T$$

Ma thèse

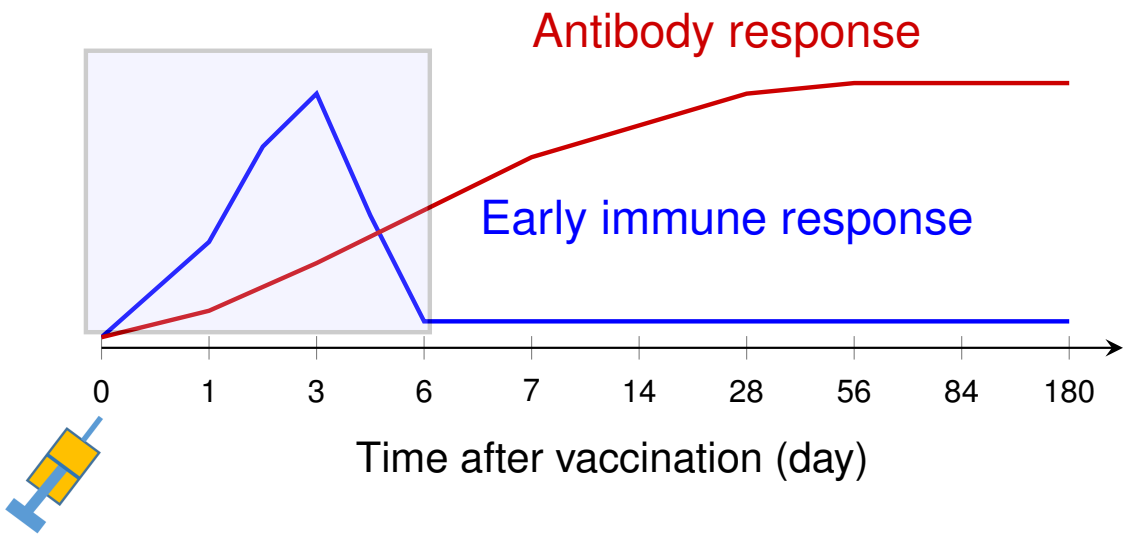
Ebola vaccine trial data set, $n = 20$ participants



Ebola vaccine trial data set, $n = 20$ participants



Ebola vaccine trial data set, $n = 20$ participants

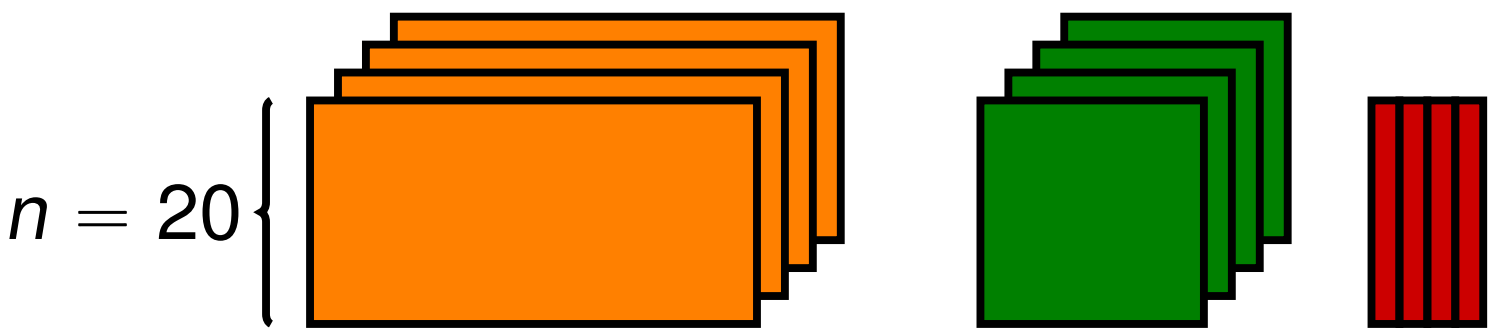


Question

How to predict the **long term antibody status** thanks to the **early immune response** ?

An High Dimensional Heterogeneous Data Set

Type	Number of	
	Variables	Measurements
Antibody response	1	4
Cellular functionnality	$\propto 10^2$	4
Transcriptome	$\propto 10^4$	4

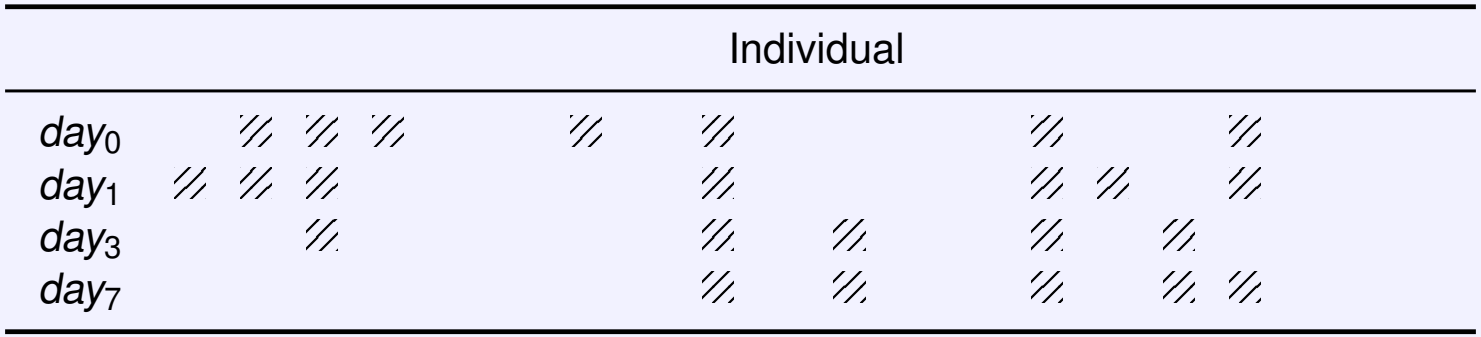


A Multiblock Description

Transcriptome : 30% of missing values (symbol : ▨)

	Individual																			
day_0		▨	▨	▨		▨	▨			▨			▨							
day_1	▨	▨	▨				▨			▨			▨	▨		▨				
day_3			▨				▨	▨		▨		▨								
day_7							▨	▨		▨		▨	▨							

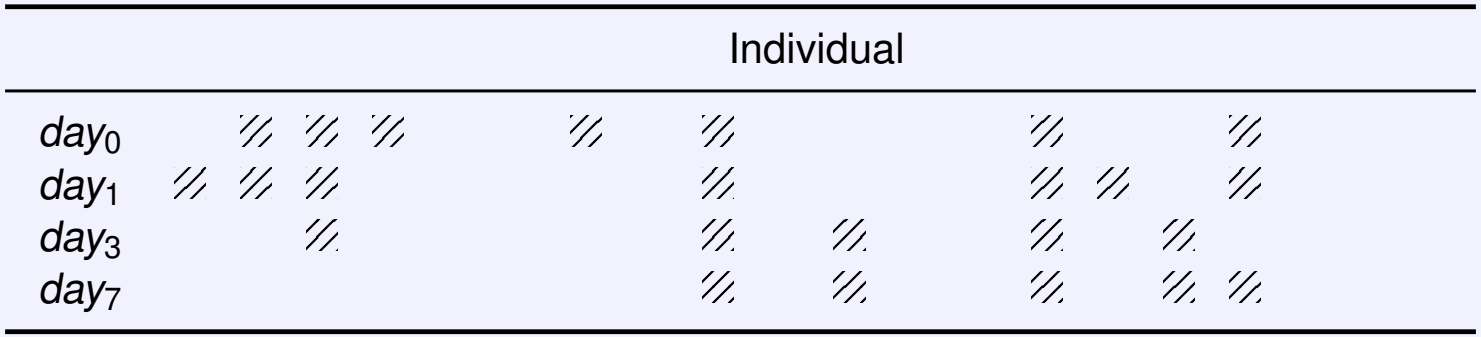
Transcriptome : 30% of missing values (symbol : ▨)



First solution

sparse PLS model [Lê Cao et al. 2008] with **nipals** imputation.
[Rechtien et al., *Cell Reports*, 2017]

Transcriptome : 30% of missing values (symbol : ▨)



First solution

sparse PLS model [Lê Cao et al. 2008] with **nipals** imputation.
[Rechtien et al., *Cell Reports*, 2017]

Research question

How to estimate missing values taking into account the response variables to be predicted.

Objectives of the thesis

Find a method which allows to deal with

- supervised problems and multivariate response ($q \geqslant 1$),
- regularization ($n << p$) and variable selection in $\mathbf{X}$ and in $\mathbf{Y}$,
- block structured covariate part,
- interpretable parameters and hyper-parameters,
- missing values per block,
- supervised imputation,
- reasonable computational time.

Objectives of the thesis

Find a method which allows to deal with

- supervised problems and multivariate response ($q \geqslant 1$),

For n individuals, explain the q variables of $\mathbf{Y}$ with the p variables of $\mathbf{X}$, and so :

- $\mathbf{X} \in \mathbb{R}^{n \times p}$: the covariate matrix,
- $\mathbf{Y} \in \mathbb{R}^{n \times q}$: the response matrix,

assumed to be standardized in the following.

Objectives of the thesis

Find a method which allows to deal with

- regularization ($n << p$) and variable selection in $\mathbf{X}$ and in $\mathbf{Y}$,

Objectives of the thesis

Find a method which allows to deal with

- block structured covariate part,

X variables can be divided in T groups of variables, with the notation

$$\begin{array}{ll} \exists T \in \mathbb{N}^\star & \mid \quad \mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_T\}, \\ \forall t \in [\![1, T]\!] & \quad \mathbf{X}_t \in \mathbb{R}^{n \times p_t} \mid p_t \in \mathbb{N}^\star, \\ & \quad p = \sum_{t=1}^{T} p_t. \end{array}$$

Objectives of the thesis

Find a method which allows to deal with

- interpretable parameters and hyper-parameters,

The regression parameters :

- block-level descriptions,
- multiblock-level descriptions,
- linear models and explained variances interpretations.

The hyper-parameters :

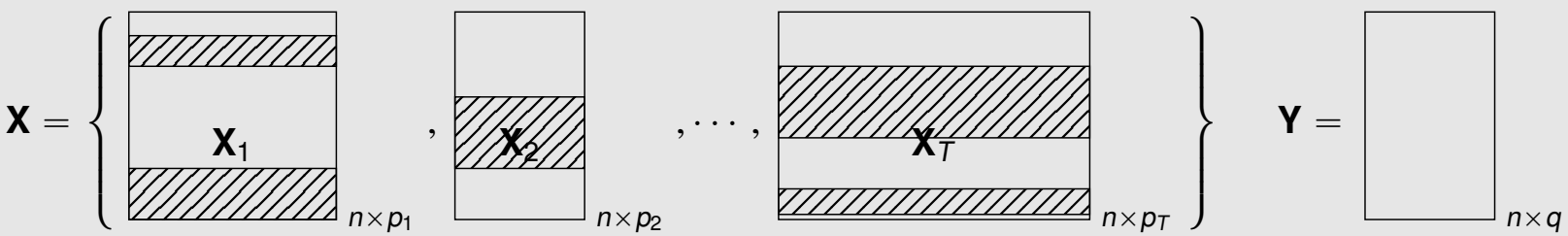
- not numerous,
- classical or easy to interpret.

Objectives of the thesis

Find a method which allows to deal with

- missing values per block,

Hatched areas represent the missing values :



Common if blocks are associated with technologies and/or time of measurements.

Objectives of the thesis

Find a method which allows to deal with

- supervised imputation,

Impute covariates associated with the response only.

Objectives of the thesis

Find a method which allows to deal with

- reasonable computational time.

Une introduction à l'apprentissage statistique
- Décomposition du risque
- Compromis biais-variance
- Le problème linéaire
- Solutions envisagées
- La régularisation

My thesis

ddsPLS, complete data
- Statistical model and estimators
- Application to a real data set

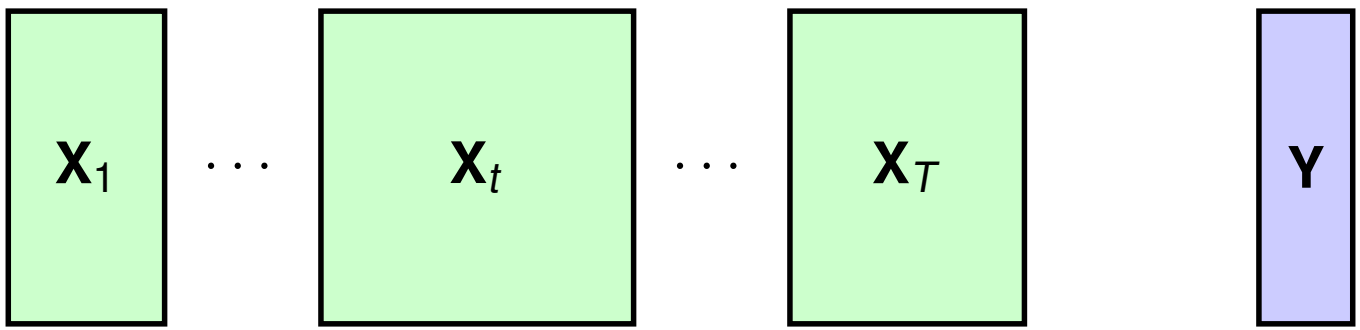
Koh-Lanta, missing values per block
- Algorithm
- Simulations
- Application to real data sets
  - Ebola data set
  - Venous thrombosis data set

Packages

Conclusions and perspectives

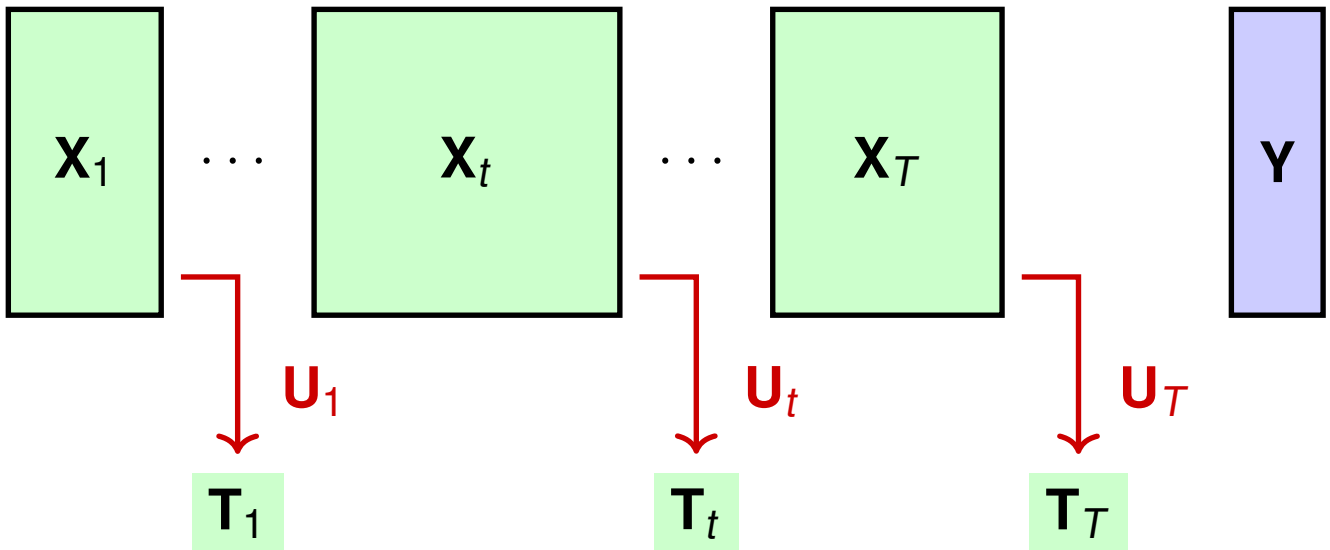
ddsPLS *data-driven sparse Partial Least Squares*

Technical points



ddsPLS *data-driven sparse Partial Least Squares*

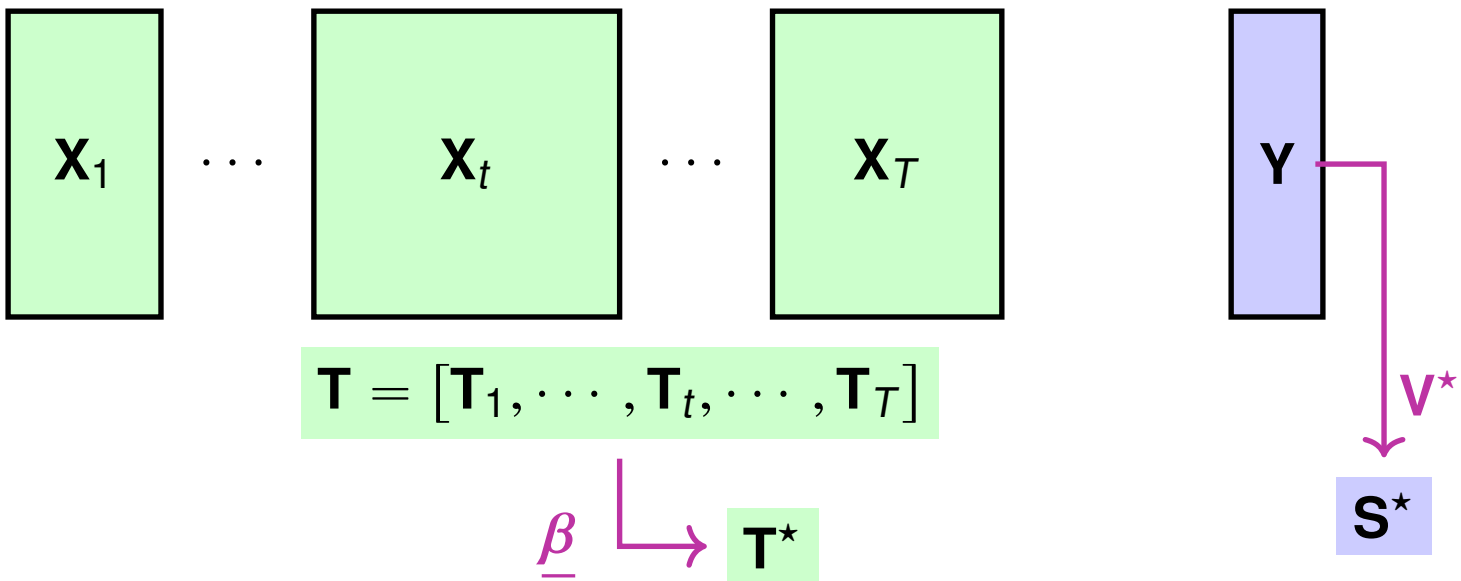
Technical points



Builds T R-dimensional sparse linear descriptions of the T blocks independently, denoting the **weights** $(\mathbf{U}_t)_{t\in[\![1,T]\!]}$.

ddsPLS *data-driven sparse Partial Least Squares*

Technical points



Builds one R-dimensional linear descriptions of the previous T descriptions, denoting the **super-weights** $(\mathbf{U}_t\underline{\beta})_{t\in[\![1,T]\!]}$ and $\mathbf{V}^\star$.

ddsPLS *data-driven sparse Partial Least Squares*

Technical points

The regularization is performed through :

- λ the minimum correlation (X_i, Y_j),
  or L_0 the maximum number of covariates, to be put in the model. See [Deshpande et Montanari 2016].

Regularization of $\mathbf{X}^T\mathbf{Y}/(n-1)$ by soft-thresholding :

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 & 0.1 \\ 0.5 & -0.2 & 0.01 \\ -0.6 & 0.1 & 0.15 \\ -0.1 & 0.03 & 0.02 \end{bmatrix}}_{\frac{\mathbf{X}^T\mathbf{Y}}{n-1}}$$

ddsPLS *data-driven sparse Partial Least Squares*

Technical points

The regularization is performed through :

- λ the minimum correlation (X_i, Y_j),
  or L_0 the maximum number of covariates, to be put in the model. See [Deshpande et Montanari 2016].

Regularization of $\mathbf{X}^T\mathbf{Y}/(n-1)$ by soft-thresholding :

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 & 0.1 \\ 0.5 & -0.2 & 0.01 \\ -0.6 & 0.1 & 0.15 \\ -0.1 & 0.03 & 0.02 \end{bmatrix}}_{\frac{\mathbf{X}^T\mathbf{Y}}{n-1}} \quad \underset{\lambda = 0.2}{\Longrightarrow} \quad \underbrace{\begin{bmatrix} \cdot & 0.7 & \cdot \\ 0.3 & \cdot & \cdot \\ -0.4 & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}_{S_\lambda\left(\frac{\mathbf{X}^T\mathbf{Y}}{n-1}\right)}$$

ddsPLS *data-driven sparse Partial Least Squares*

Technical points

The regularization is performed through :

- λ the minimum correlation (X_i, Y_j),
  or L_0 the maximum number of covariates, to be put in the model. See [Deshpande et Montanari 2016].

Regularization of $\mathbf{X}^T\mathbf{Y}/(n-1)$ by soft-thresholding :

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 & 0.1 \\ 0.5 & -0.2 & 0.01 \\ -0.6 & 0.1 & 0.15 \\ -0.1 & 0.03 & 0.02 \end{bmatrix}}_{\frac{\mathbf{X}^T\mathbf{Y}}{n-1}} \xrightarrow[\lambda = 0.2]{} \underbrace{\begin{bmatrix} \cdot & 0.7 & \cdot \\ 0.3 & \cdot & \cdot \\ -0.4 & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}_{S_\lambda\left(\frac{\mathbf{X}^T\mathbf{Y}}{n-1}\right)} \xrightarrow[\text{SVD}]{} \begin{cases} \mathbf{u}^{(1)} = [1 & \cdot & \cdot & \cdot]^T \\ \mathbf{u}^{(2)} = [\cdot & -0.6 & 0.8 & \cdot]^T \\ \mathbf{u}^{(3)} = [\cdot & 0.8 & 0.6 & \cdot]^T \\ \mathbf{u}^{(4)} = [\cdot & \cdot & \cdot & 1]^T \end{cases}.$$

ddsPLS *data-driven sparse Partial Least Squares*

Technical points

The regularization is performed through :

- λ the minimum correlation (X_i, Y_j),
  or L_0 the maximum number of covariates, to be put in the model. See [Deshpande et Montanari 2016].

Regularization of $\mathbf{X}^T\mathbf{Y}/(n-1)$ by soft-thresholding :

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 & 0.1 \\ 0.5 & -0.2 & 0.01 \\ -0.6 & 0.1 & 0.15 \\ -0.1 & 0.03 & 0.02 \end{bmatrix}}_{\frac{\mathbf{X}^T\mathbf{Y}}{n-1}} \xrightarrow[\lambda = 0.2]{} \underbrace{\begin{bmatrix} \cdot & 0.7 & \cdot \\ 0.3 & \cdot & \cdot \\ -0.4 & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}_{S_\lambda\left(\frac{\mathbf{X}^T\mathbf{Y}}{n-1}\right)} \xrightarrow[\text{SVD}]{} \begin{cases} \mathbf{u}^{(1)} = [1 & \cdot & \cdot & \cdot]^T \\ \mathbf{u}^{(2)} = [\cdot & -0.6 & 0.8 & \cdot]^T \\ \mathbf{u}^{(3)} = [\cdot & 0.8 & 0.6 & \cdot]^T \end{cases}.$$

- R the number of components, here $R = 3$.

(λ, R) or (L_0, R) optimized through cross-validation.

Recall : There are no missing values here.

Statistical model

Single observation level model

$\forall t \in [\![1, T]\!]$, $w_t \in \mathbb{R}^R$ is a latency random vector associated with $y \in \mathbb{R}^q$ and $x_t \in \mathbb{R}^{p_t}$ such as

$$\begin{cases} \forall t \in [\![1, T]\!], & w_t = x_t \mathbf{U}_t + \epsilon_t \\ \text{and} & y = \sum_{t=1}^{T} w_t \mathbf{Q}_t + \epsilon_y, \end{cases} \tag{8}$$

where $\epsilon_t \sim \mathcal{N}\left(\mathbf{0}, \sigma_t^2 \mathbb{I}_R\right)$ and $\epsilon_y \sim \mathcal{N}\left(\mathbf{0}, \sigma_y^2 \mathbb{I}_q\right)$ are residuals.

Statistical model

Train data set level model

$\mathcal{D}_n$ a train data set of n independent observations, under matrix notations, it becomes

$$\begin{cases} \forall t \in [\![1, T]\!], & \mathbf{W}_t = \mathbf{X}_t\mathbf{U}_t + \mathbf{E}_t \\ \text{and} & \mathbf{Y} = \sum_{t=1}^{T} \mathbf{W}_t\mathbf{Q}_t + \mathbf{E}_y. \end{cases} \tag{8}$$

Regression model

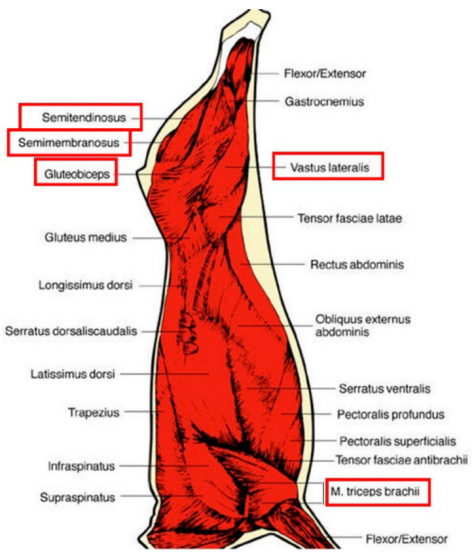
Model (1) directly implies the regression model

$$\mathbf{Y} = \sum_{t=1}^{T} \mathbf{X}_t\mathbf{B}_t + \mathbf{F}, \tag{9}$$

where $\mathbf{B}_t = \mathbf{U}\mathbf{Q}_t'$ and $\mathbf{F}$ is a residual matrix which is a function of

Application to a real data set

Application to a complete data set : selection of biomarkers the most predictive of tenderness



Data set structure

$q = 5, T = 5, n = 10,$

$p_1 = \cdots = p_5 = 20.$

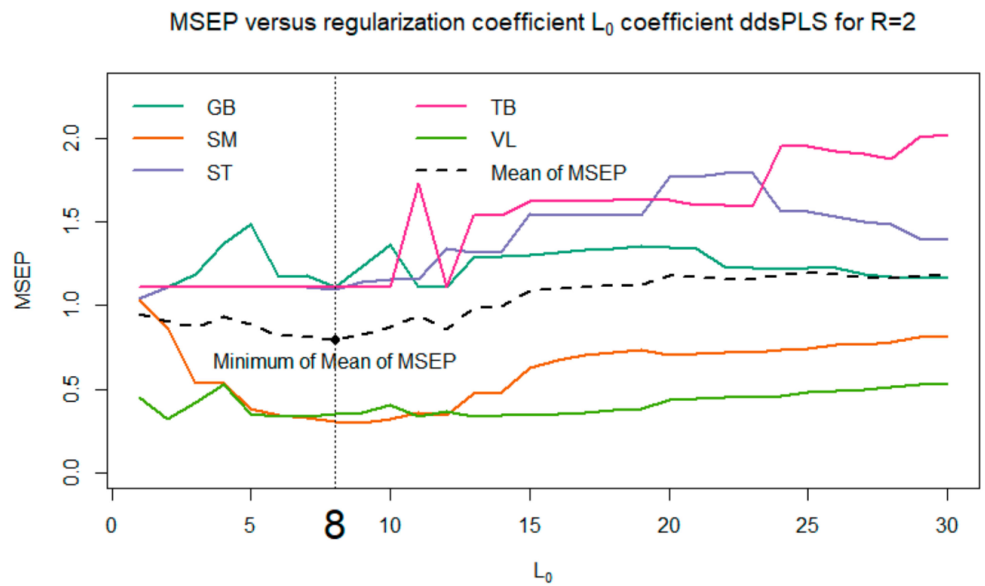
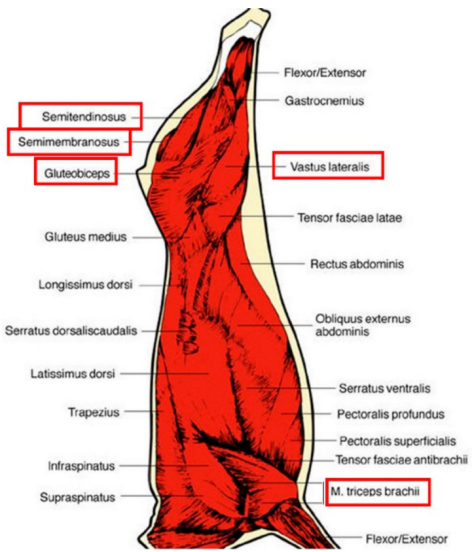


FIGURE – Leave-one-out cross-validation results

Application to a complete data set : selection of biomarkers the most predictive of tenderness



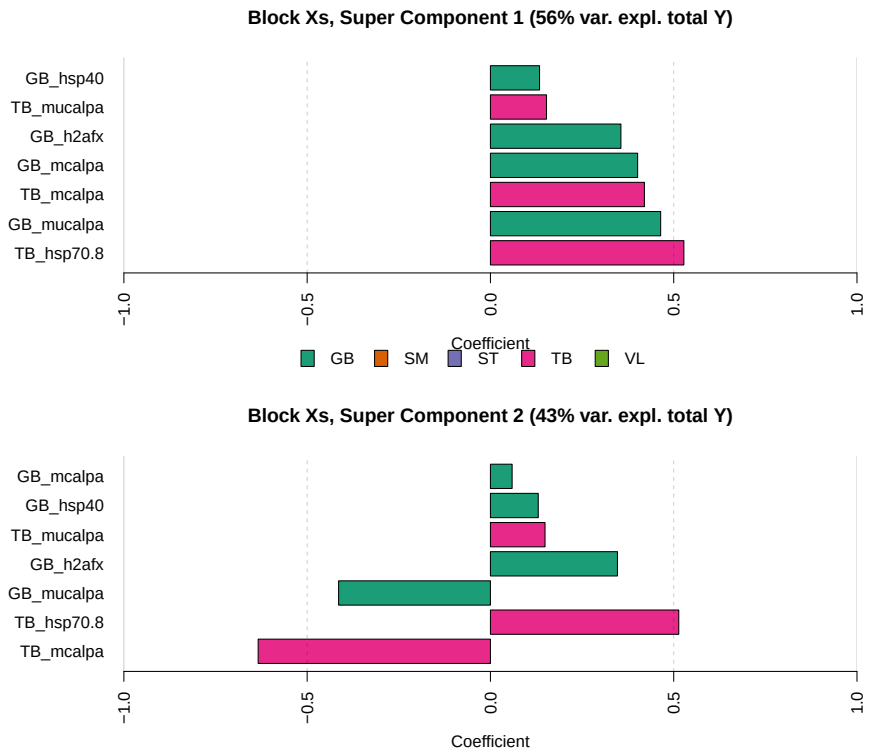


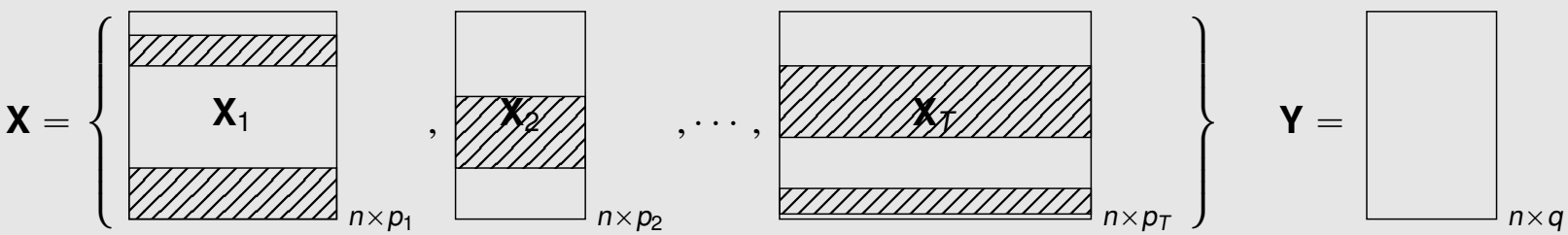
FIGURE – Model super-weights

[LORENZO et al., *Foods*, 2019]

Une introduction à l'apprentissage statistique
Décomposition du risque
Compromis biais-variance
Le problème linéaire
Solutions envisagées
La régularisation
My thesis
ddsPLS, complete data
Statistical model and estimators
Application to a real data set
Koh-Lanta, missing values per block
Algorithm
Simulations
Application to real data sets
Ebola data set
Venous thrombosis data set
Packages
Conclusions and perspectives

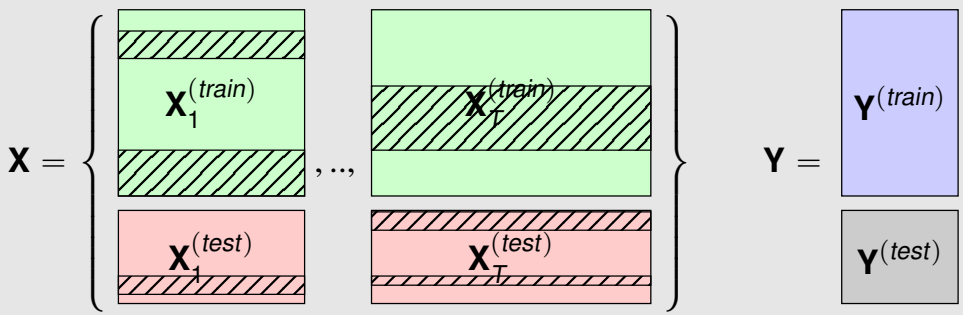
Koh-Lanta : impute missing values per block in supervised context

Recall, the hatched areas represent the missing values :



Koh-Lanta : impute missing values per block in supervised context

In fact : missing values can occur in $\mathbf{X}^{(train)}$ and in $\mathbf{X}^{(test)}$:



Koh-Lanta : a two step algorithm

The Tribe Stage, imputes the $\mathbf{X}^{(train)}$ part :

- EM typed algorithm, estimation of the T structures, of the missing values and of the overall structure.

The Reunification Stage, imputes the $\mathbf{X}^{(test)}$ part :

- no alternating algorithm $\Longrightarrow$ No convergence criterion.

Notations

- "$\mathcal{V}_t^\star$" is the list of selected variables in $\mathbf{X}_t^\star$.

Notations

- "$\mathcal{V}_t^\star$" is the list of selected variables in $\mathbf{X}_t^\star$.
- "$\mathbf{X}_t^{\star,\mathcal{V}_t^\star}$" the columns of $\mathbf{X}_t^\star$ corresponding to $\mathcal{V}_t^\star$.

Notations

- "$\mathcal{V}_t^\star$" is the list of selected variables in $\mathbf{X}_t^\star$.
- "$\mathbf{X}_t^{\star,\mathcal{V}_t^\star}$" the columns of $\mathbf{X}_t^\star$ corresponding to $\mathcal{V}_t^\star$.
- "$\mathcal{M} = \mathbf{ddsPLS}(x = \bullet_1, y = \bullet_2, \lambda, R)$", a ddsPLS model with "$\bullet_1$" as covariates, "$\bullet_2$" as responses.

Notations

- "$\mathcal{V}_t^\star$" is the list of selected variables in $\mathbf{X}_t^\star$.
- "$\mathbf{X}_t^{\star,\mathcal{V}_t^\star}$" the columns of $\mathbf{X}_t^\star$ corresponding to $\mathcal{V}_t^\star$.
- "$\mathcal{M} = \textbf{ddsPLS}(x = \bullet_1, y = \bullet_2, \lambda, R)$", a ddsPLS model with
  "$\bullet_1$" as covariates, "$\bullet_2$" as responses.
- "$\textbf{predict}(\mathcal{M}, x = \bullet)$" predicts values for
  a new sample "$\bullet$", with model "$\mathcal{M}$".

Notations

- "$\mathcal{V}_t^\star$" is the list of selected variables in $\mathbf{X}_t^\star$.
- "$\mathbf{X}_t^{\star,\mathcal{V}_t^\star}$" the columns of $\mathbf{X}_t^\star$ corresponding to $\mathcal{V}_t^\star$.
- "$\mathcal{M} = \mathbf{ddsPLS}(x = \bullet_1, y = \bullet_2, \lambda, R)$", a ddsPLS model with
  "$\bullet_1$" as covariates, "$\bullet_2$" as responses.
- "$\mathbf{predict}(\mathcal{M}, x = \bullet)$" predicts values for
  a new sample "$\bullet$", with model "$\mathcal{M}$".

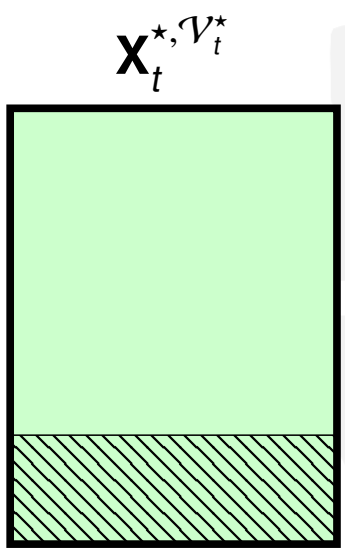
Initialisation

$$\mathcal{V}_t^\star = \left\{\text{indices of non null rows in } S_\lambda\left(\text{cor}\left(\mathbf{X}_t^o, \mathbf{Y}\right)\right)\right\},$$

$\mathbf{X}_t^o$: $\mathbf{X}_t$ imputed to the mean.

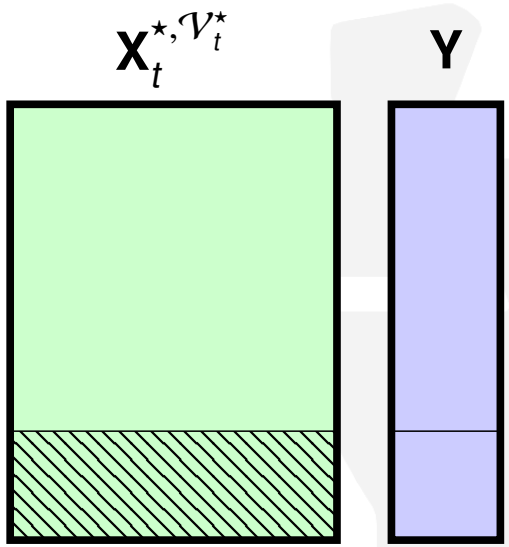
The Tribe Stage

♣ $\forall t \in [\![1, T]\!]$:



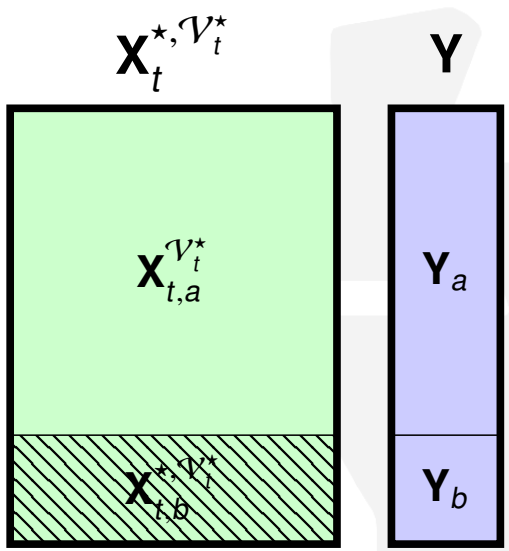
The Tribe Stage

♣ $\forall t \in [\![1, T]\!]$:



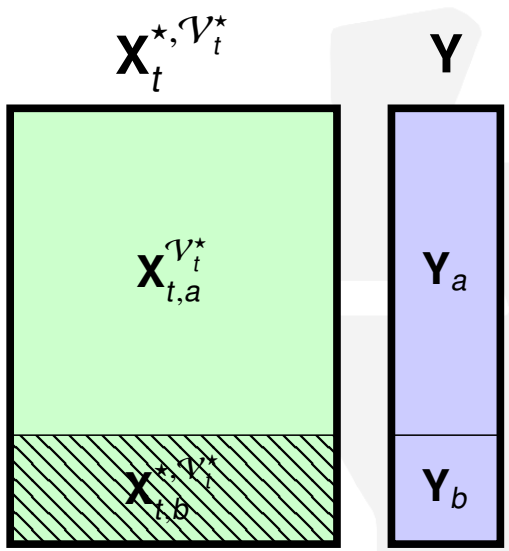
The Tribe Stage

♣ $\forall t \in [\![1, T]\!]$:



The Tribe Stage

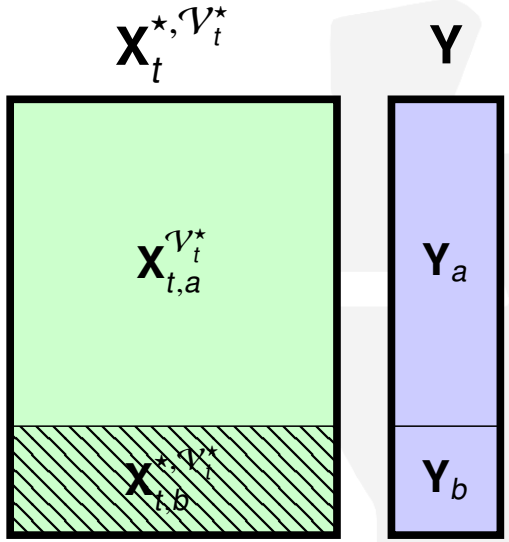
♣ $\forall t \in [\![1, T]\!]$:



$$\mathcal{M}_t^\star = \textbf{ddsPLS}\left(x = \mathbf{Y}_a, y = \mathbf{X}_{t,a}^{\mathcal{V}_t^\star}, \lambda, R\right)$$

The Tribe Stage

♣ $\forall t \in [\![1, T]\!]$:

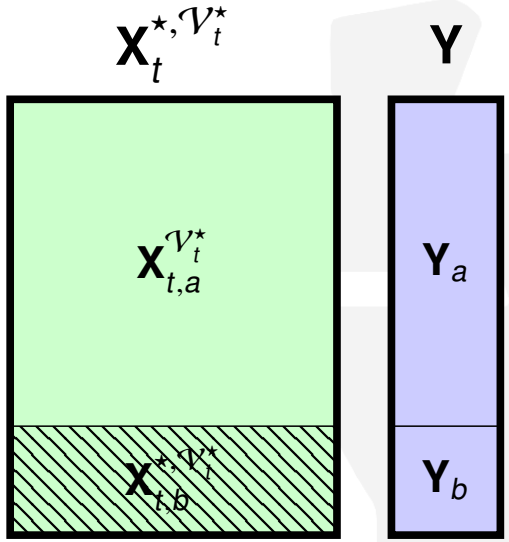


$$\mathcal{M}_t^\star = \textbf{ddsPLS}\left(x = \mathbf{Y}_a, y = \mathbf{X}_{t,a}^{\mathcal{V}_t^\star}, \lambda, R\right)$$

$$\widehat{\mathbf{X}_{t,b}^{\mathcal{V}_t^\star}} = \textbf{predict}\left(\mathcal{M}_t^\star, \mathbf{Y}_b\right)$$

The Tribe Stage

- $\forall t \in [\![1, T]\!]$:



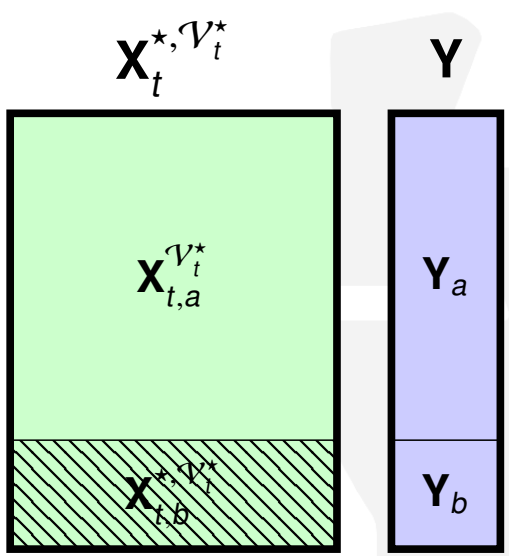
$$\mathcal{M}_t^\star = \textbf{ddsPLS}\left(x = \mathbf{Y}_a, y = \mathbf{X}_{t,a}^{\mathcal{V}_t^\star}, \lambda, R\right)$$

$$\widehat{\mathbf{X}_{t,b}^{\mathcal{V}_t^\star}} = \textbf{predict}\left(\mathcal{M}_t^\star, \mathbf{Y}_b\right)$$

- $\mathcal{M}^\star = \textbf{ddsPLS}(x = \left\{\mathbf{X}_t^\star\right\}_{t\in[\![1,T]\!]}, y = \mathbf{Y}, \lambda, R)$
  Update $\mathcal{V}_t^\star$ from $\mathcal{M}^\star$

The Tribe Stage

- $\forall t \in [\![1, T]\!]$:



$$\mathcal{M}_t^\star = \textbf{ddsPLS}\left(x = \mathbf{Y}_a, y = \mathbf{X}_{t,a}^{\mathcal{V}_t^\star}, \lambda, R\right)$$

$$\widehat{\mathbf{X}_{t,b}^{\mathcal{V}_t^\star}} = \textbf{predict}\left(\mathcal{M}_t^\star, \mathbf{Y}_b\right)$$

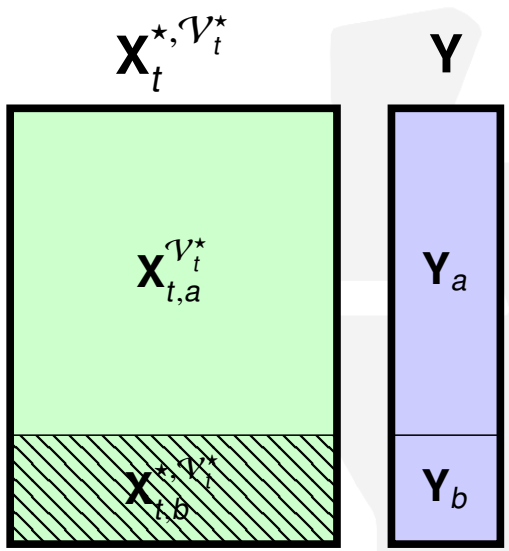
- $\mathcal{M}^\star = \textbf{ddsPLS}(x = \left\{\mathbf{X}_t^\star\right\}_{t\in[\![1,T]\!]}, y = \mathbf{Y}, \lambda, R)$

  Update $\mathcal{V}_t^\star$ from $\mathcal{M}^\star$

The Tribe Stage

Until convergence of $\mathcal{V}^\star$

- $\forall t \in [\![1, T]\!]$:



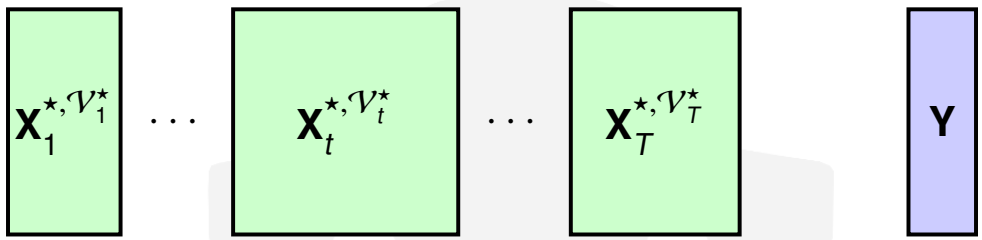
$$\mathcal{M}_t^\star = \textbf{ddsPLS}\left(x = \mathbf{Y}_a, y = \mathbf{X}_{t,a}^{\mathcal{V}_t^\star}, \lambda, R\right)$$

$$\widehat{\mathbf{X}_{t,b}^{\mathcal{V}_t^\star}} = \textbf{predict}\left(\mathcal{M}_t^\star, \mathbf{Y}_b\right)$$

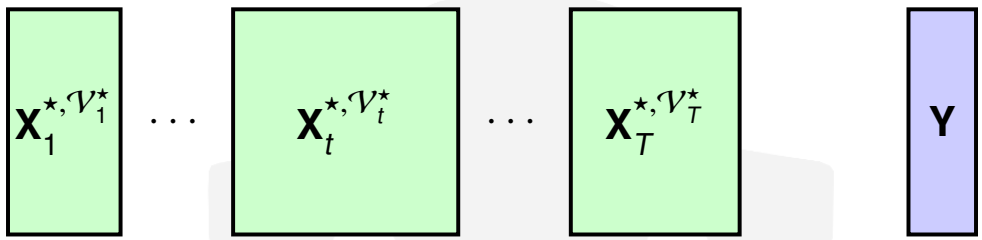
- $\mathcal{M}^\star = \textbf{ddsPLS}(x = \left\{\mathbf{X}_t^\star\right\}_{t\in[\![1,T]\!]}, y = \mathbf{Y}, \lambda, R)$

  Update $\mathcal{V}_t^\star$ from $\mathcal{M}^\star$

The Reunification Stage

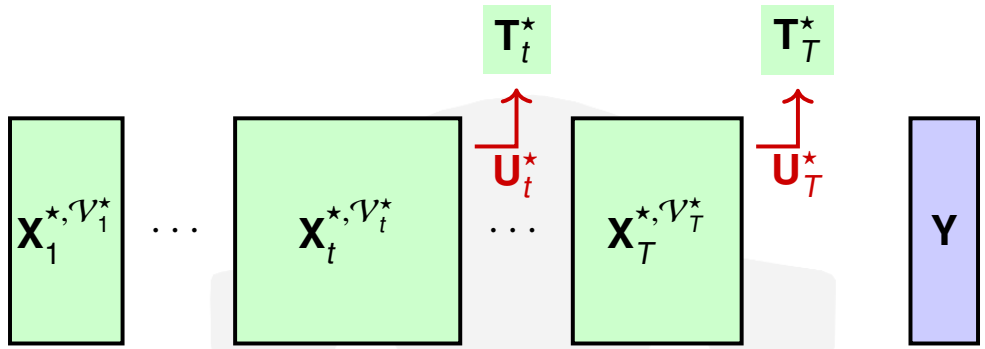


The Reunification Stage

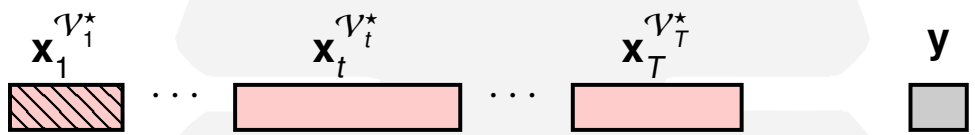


- $\mathbf{x}$ test observation. Missing values in block 1 :

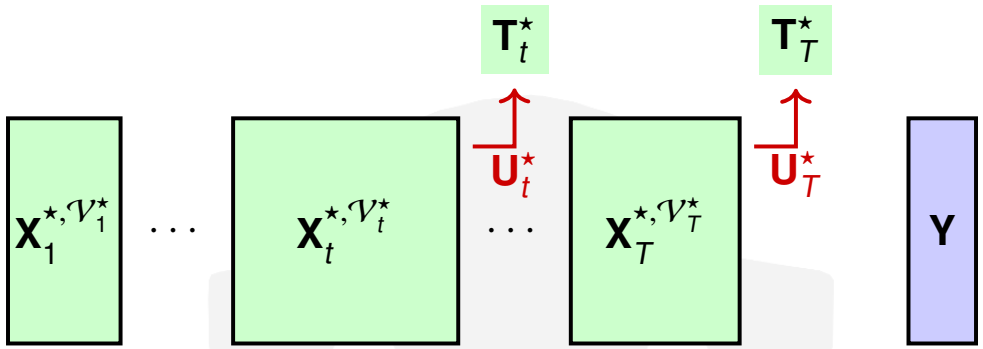
The Reunification Stage



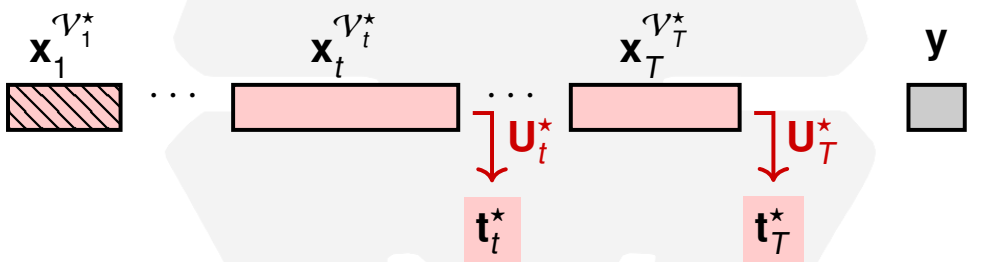
- $\mathbf{x}$ test observation. Missing values in block 1 :



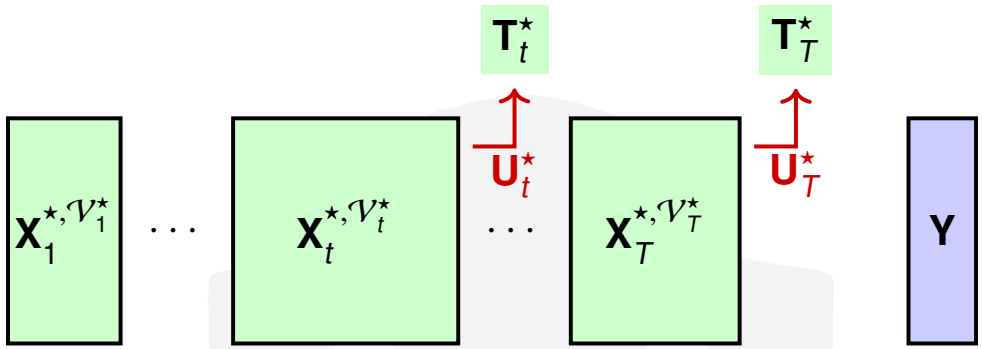
The Reunification Stage



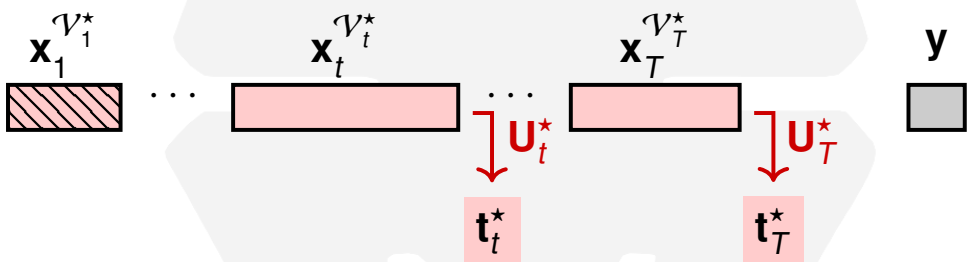
- $\mathbf{x}$ test observation. Missing values in block 1 :



The Reunification Stage

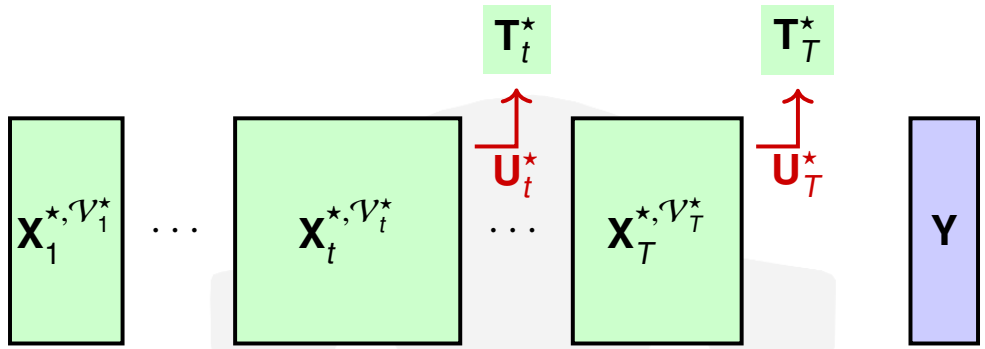


♣ $\mathbf{x}$ test observation. Missing values in block 1 :

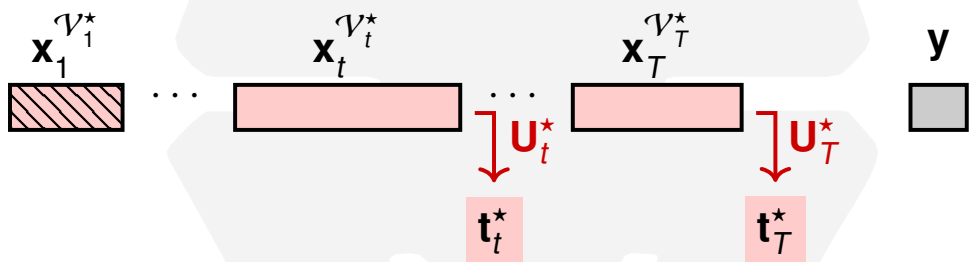


$$\mathcal{M}^\star(\mathbf{x}) = \mathbf{ddsPLS}(x = \left[\mathbf{T}_t^\star, \cdots, \mathbf{T}_T^\star\right], y = \mathbf{X}_1^{\star,\mathcal{V}_1^\star}, \lambda, R)$$

The Reunification Stage



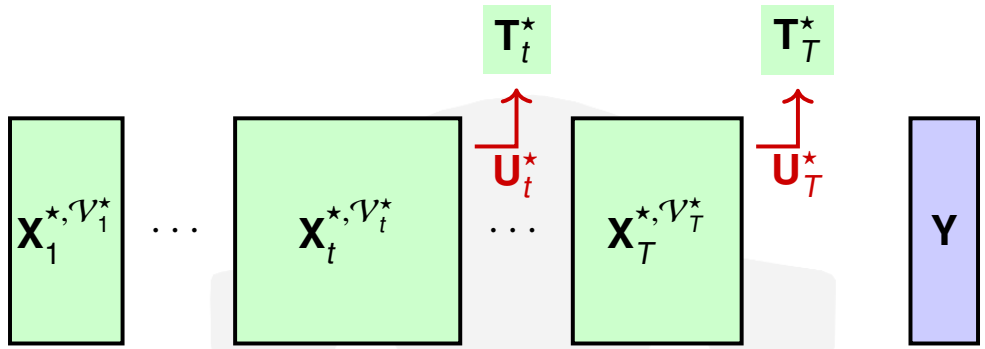
♣ $\mathbf{x}$ test observation. Missing values in block 1 :



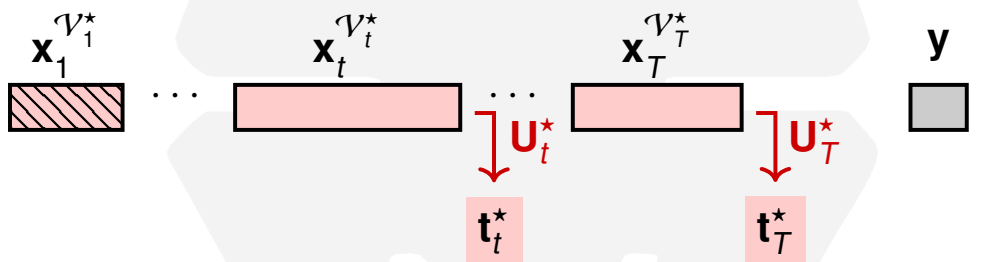
$$\mathcal{M}^\star(\mathbf{x}) = \mathbf{ddsPLS}(x = \left[\mathbf{T}_t^\star, \cdots, \mathbf{T}_T^\star\right], y = \mathbf{X}_1^{\star,\mathcal{V}_1^\star}, \lambda, R)$$

$$\widehat{\mathbf{x}_1^{\mathcal{V}_1^\star}} = \mathbf{predict}\left(\mathcal{M}^\star(\mathbf{x}), \left[\mathbf{t}_t^\star, \cdots, \mathbf{t}_T^\star\right]\right)$$

The Reunification Stage



♣ $\mathbf{x}$ test observation. Missing values in block 1 :



$$\mathcal{M}^\star(\mathbf{x}) = \mathbf{ddsPLS}(x = \left[\mathbf{T}_t^\star, \cdots, \mathbf{T}_T^\star\right], y = \mathbf{X}_1^{\star,\mathcal{V}_1^\star}, \lambda, R)$$

$$\widehat{\mathbf{x}_1^{\mathcal{V}_1^\star}} = \mathbf{predict}\left(\mathcal{M}^\star(\mathbf{x}), \left[\mathbf{t}_t^\star, \cdots, \mathbf{t}_T^\star\right]\right) \qquad \widehat{\mathbf{y}} = \mathbf{predict}\left(\mathcal{M}^\star, \left[\widehat{\mathbf{x}_1}, \cdots, \mathbf{x}_T\right]\right)$$

Simulations

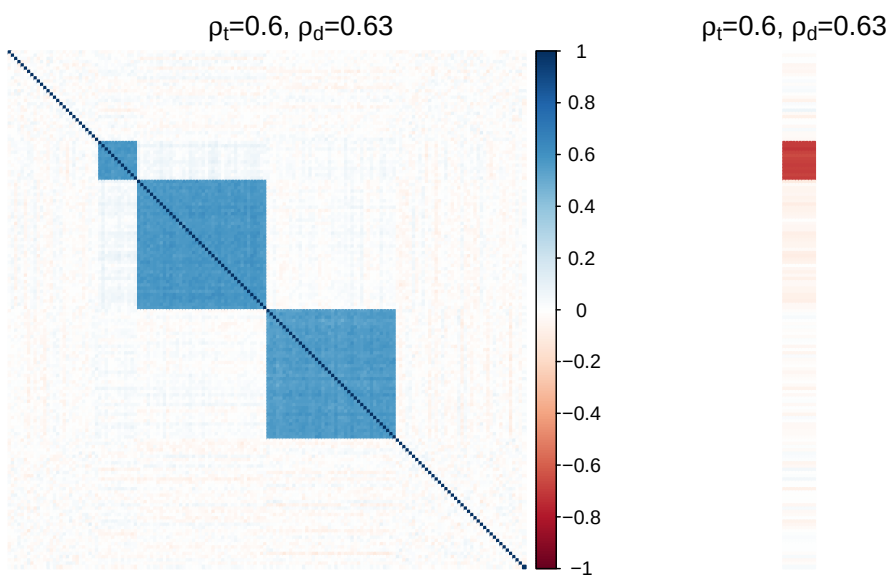
Simulation design

$p_t = 160$ variables. 3 groups of correlated variables with ρ_d :

- $p_t^{(1)} \sim \mathcal{U}([\![0, 40]\!])$,associated with the response.
- $p_t^{(2)} = p_t^{(3)} = 40$, not associated with the response.

$n = 1000$ $\quad \mathbf{X}_t \quad \mathbf{Y}$

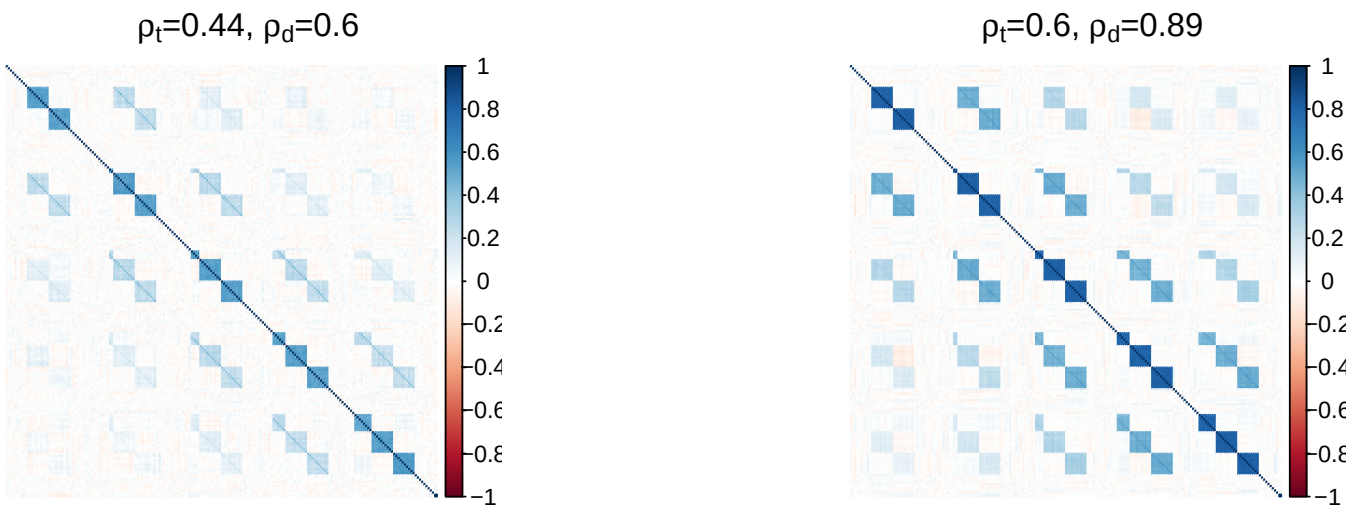
$$p_t^{(1)} \sim \mathcal{U}([\![1, 40]\!]) \quad p_t^{(2)} = 40 \quad p_t^{(3)} = 40$$



Simulation design

- ρ_d controls the intra-block correlation.
- ρ_t controls the inter-block correlation.

Examples for $T = 5$, $n = 1000$, $p_t^{(1)} = 8(t-1)$



(a) Inter-block correlation, effect of ρ_t. (b) Intra-block correlation, effect of ρ_d.

Baseline methods & question

2 step methods :

- Imputation : **imputeMFA**(missMDA) [Husson et Josse 2013], **softImpute** [Hastie et al. 2015], **Mean**, **nipals** (mixOmics solution),
- Prediction : **ddsPLS**, **Lasso** classical **sPLS** (for **nipals** imputation).

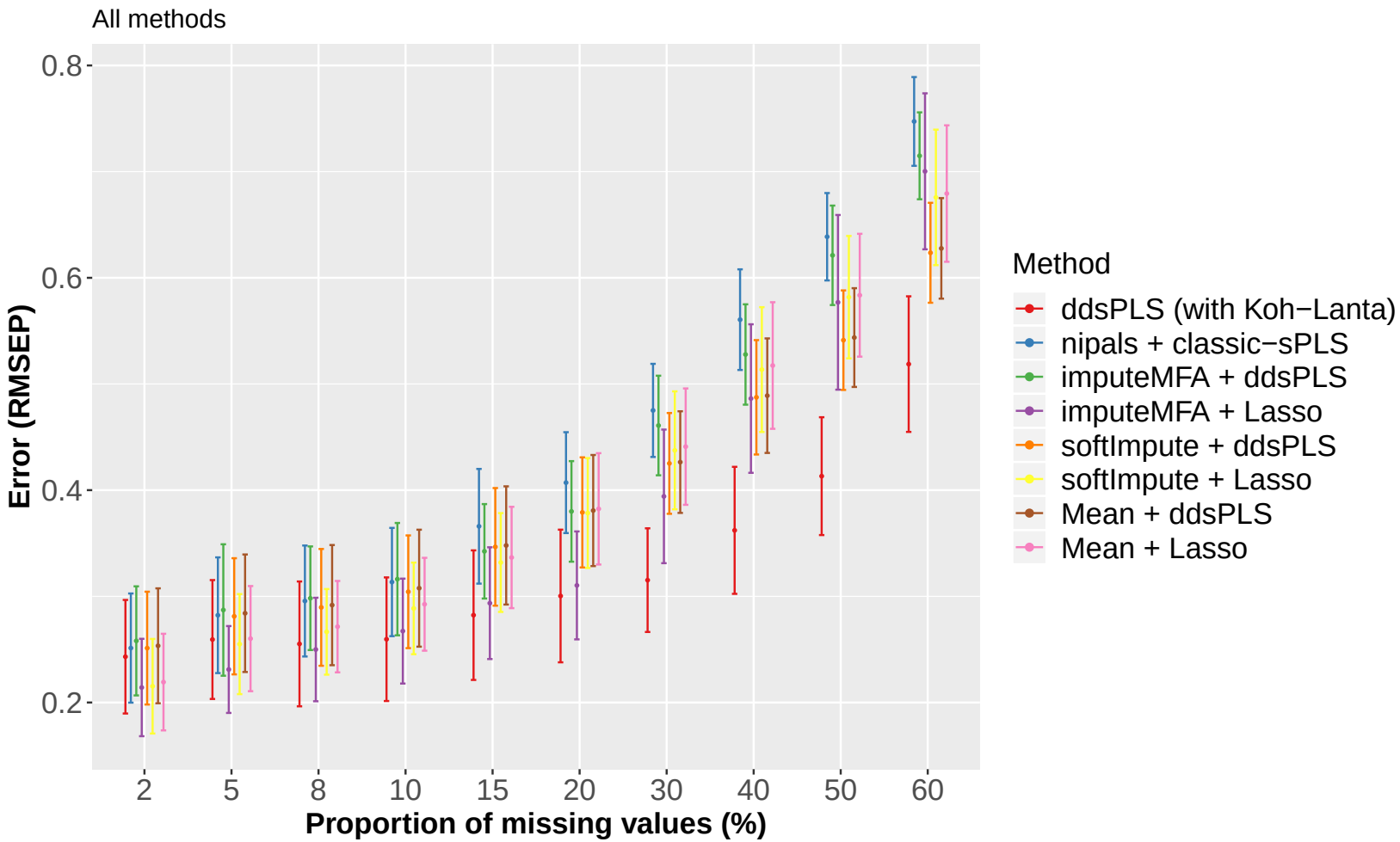
All-in-One method : [**che2018recurrent**] : classification problems, recurrent neural networks, huge n.

Simulation questions

- Robustness to increasing number of missing values ?
- Robustness to low n and $n << p$?
- Robustness to low inter/intra-block correlations ?

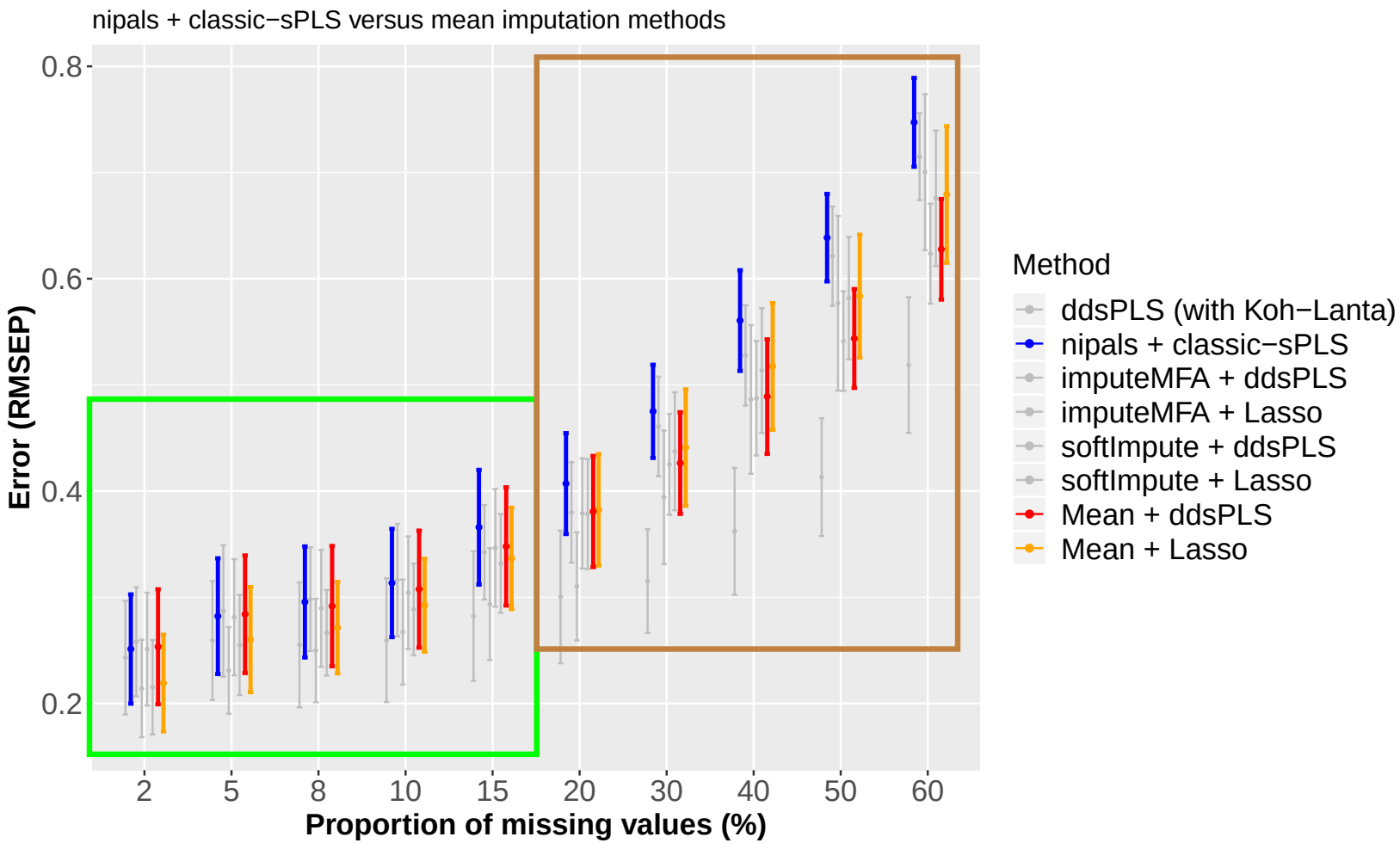
An example of simulations with new function **imputeMFA**

$n = 100, \rho_t = \rho_d = 0.9.$



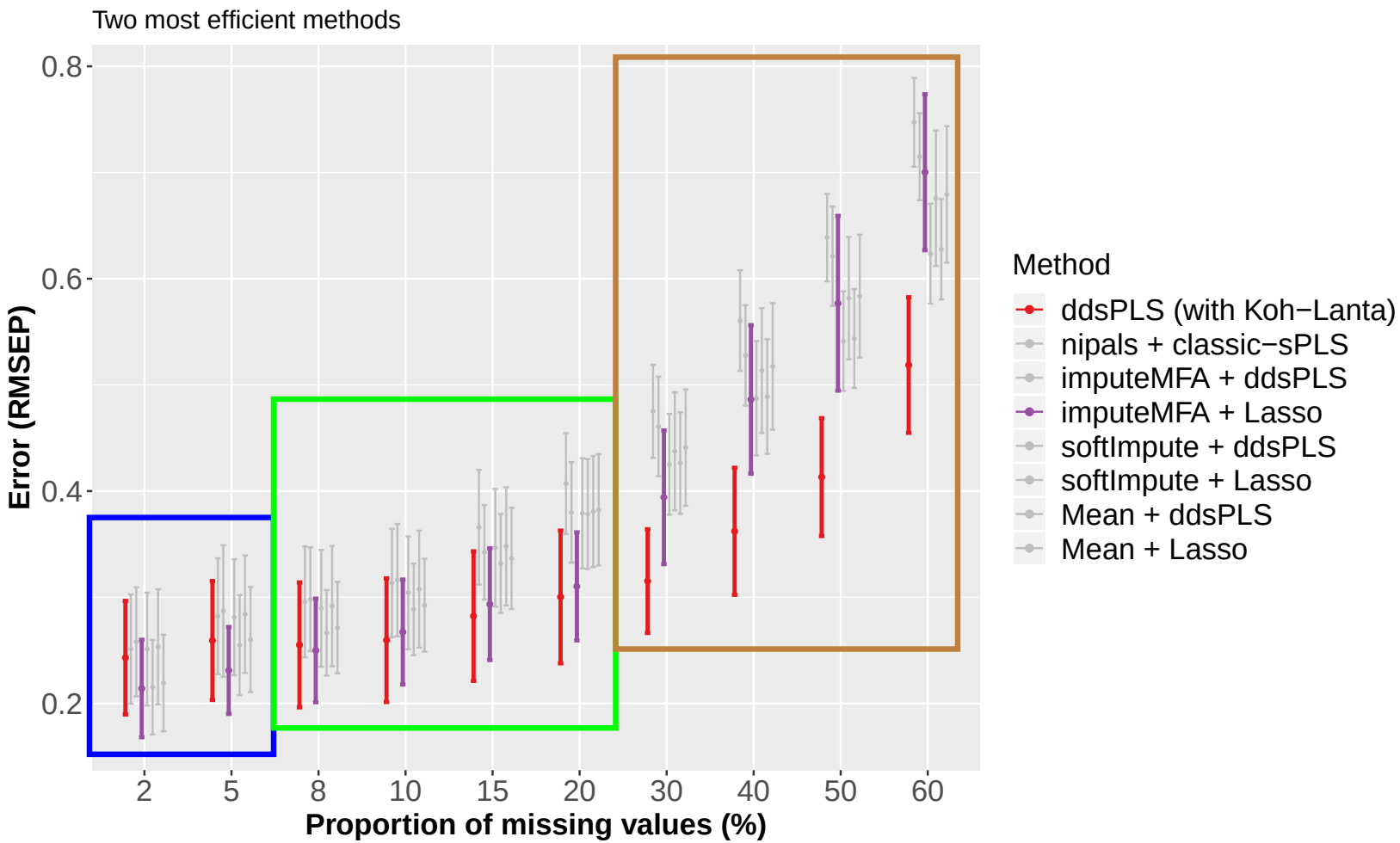
An example of simulations with new function **imputeMFA**

$n = 100, \rho_t = \rho_d = 0.9.$



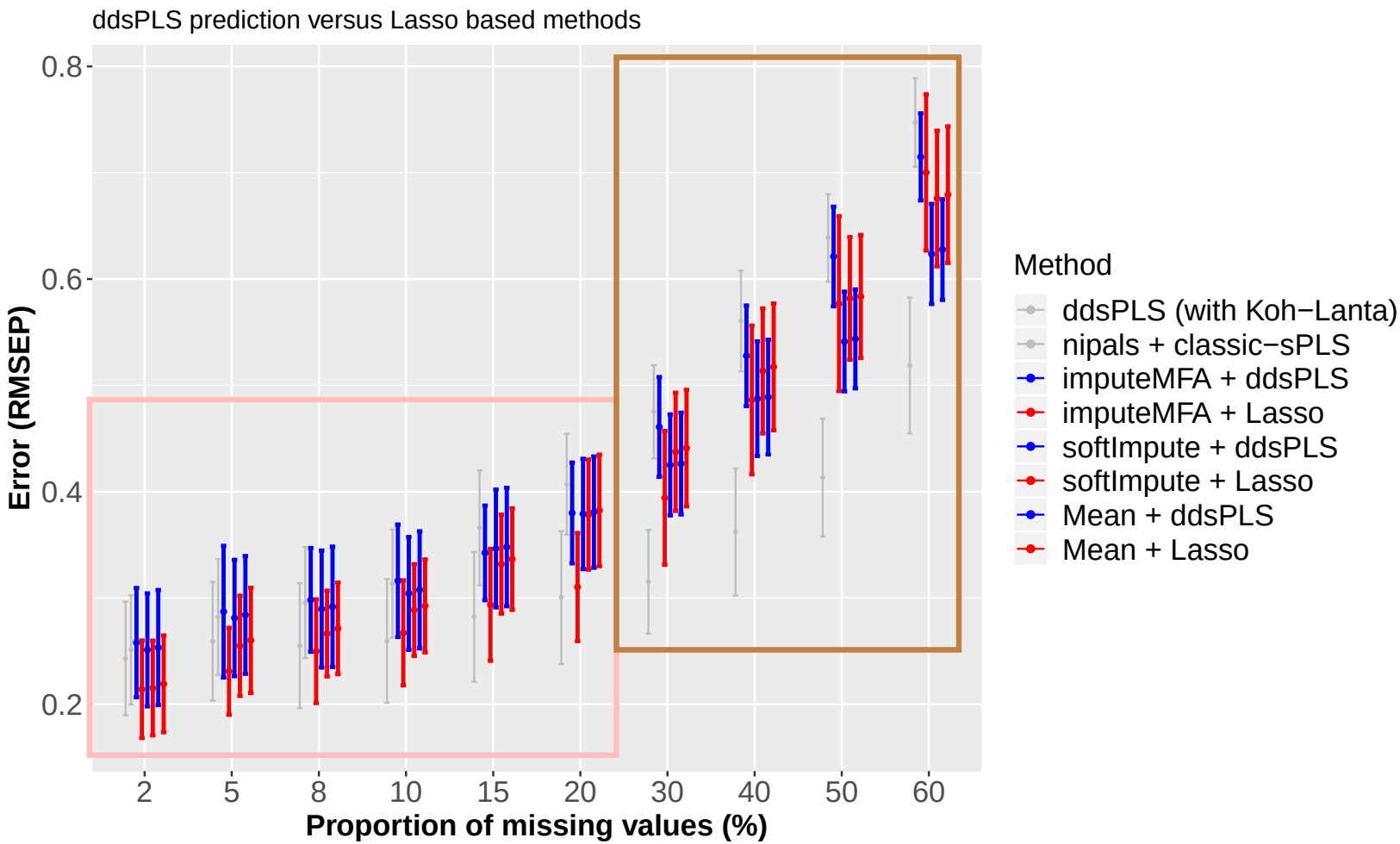
An example of simulations with new function **imputeMFA**

$n = 100, \rho_t = \rho_d = 0.9.$



An example of simulations with new function **imputeMFA**

$n = 100, \rho_t = \rho_d = 0.9.$



Application to real data sets

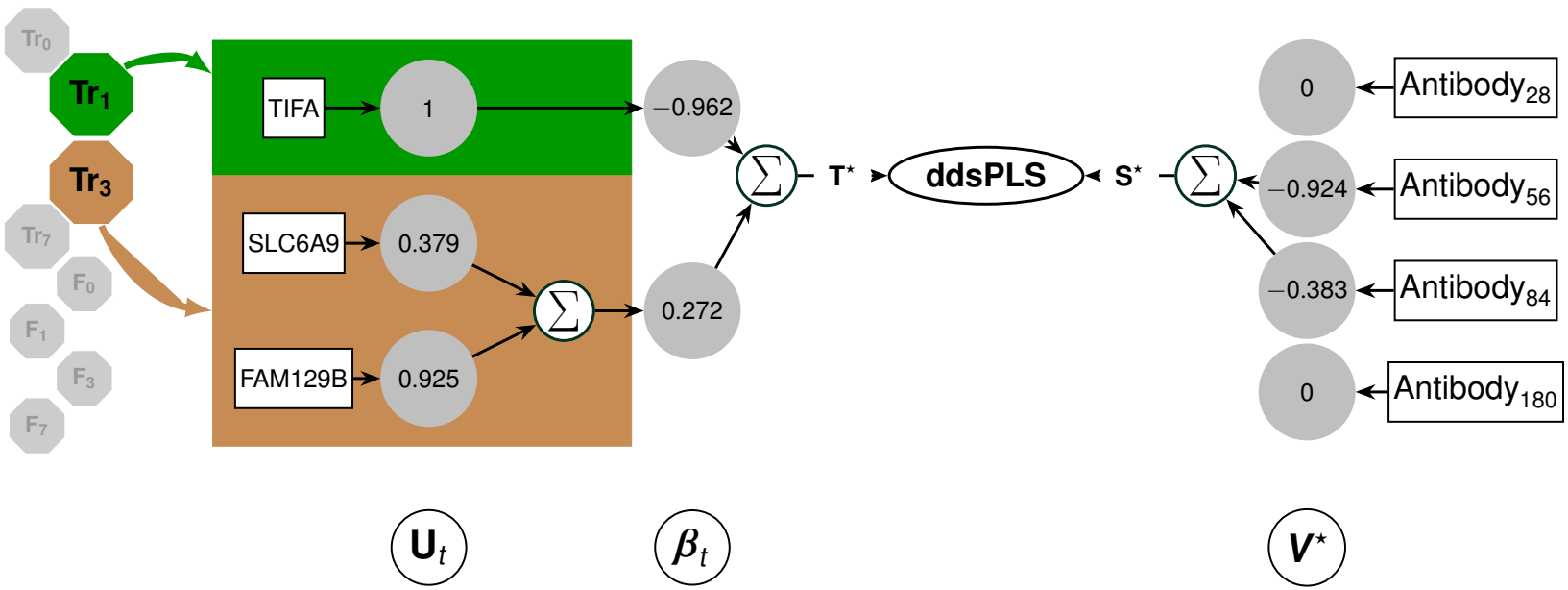
Application to the Ebola data-set

Comparaison Koh-Lanta/Mean imputation for ddsPLS model

Leave-one-out results	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

Application to the Ebola data-set

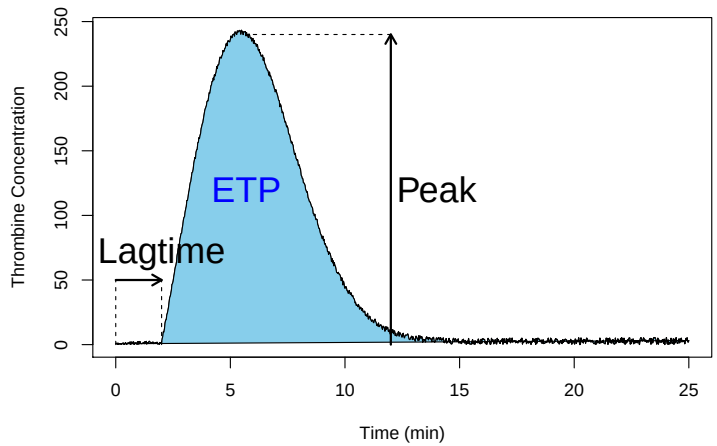
Final model : ddsPLS with Koh-Lanta



Venous thrombosis data set

3 main biomarkers

- Lagtime
  → *Delay*,
- Peak
  → *Maximum value*,
- ETP
  → *Area under the curve*.



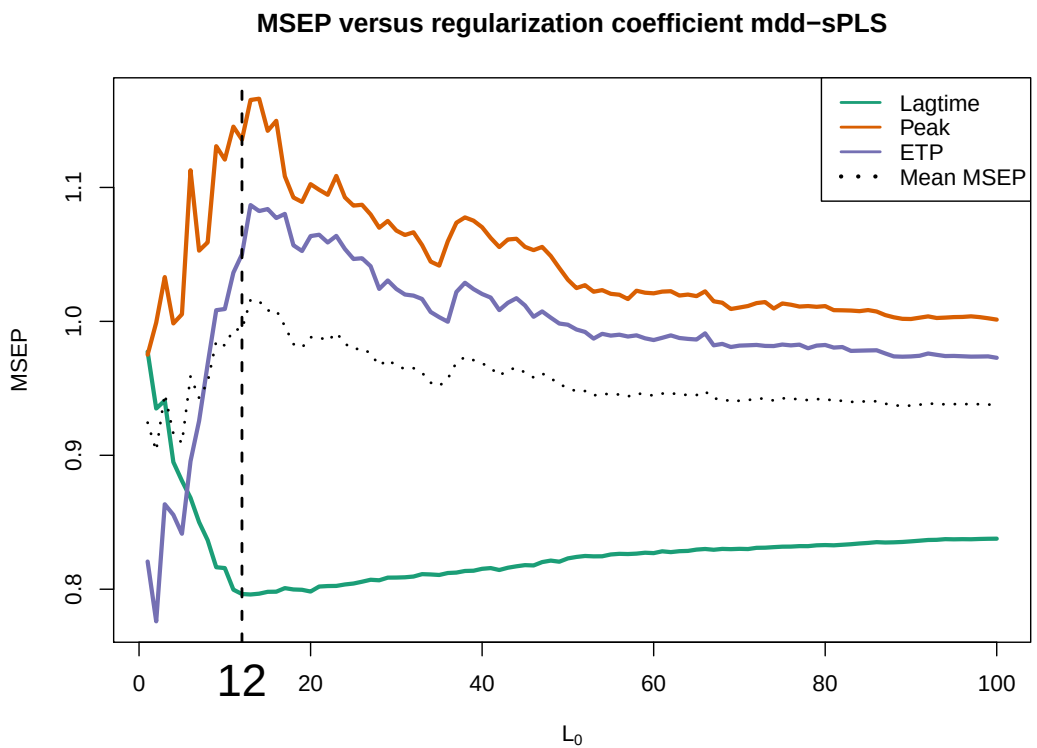
Data set structure

$n = 696$, $q = 3$, $T = 5$, $p_1 = 384$, $p_2 = 3000$, $p_3 = 1$, $p_4 = p_5 = 3$.
Biggest challenge :

- $\approx$ 68% of missing values of the DNA methylation, $t = 2$

Venous thrombosis data set

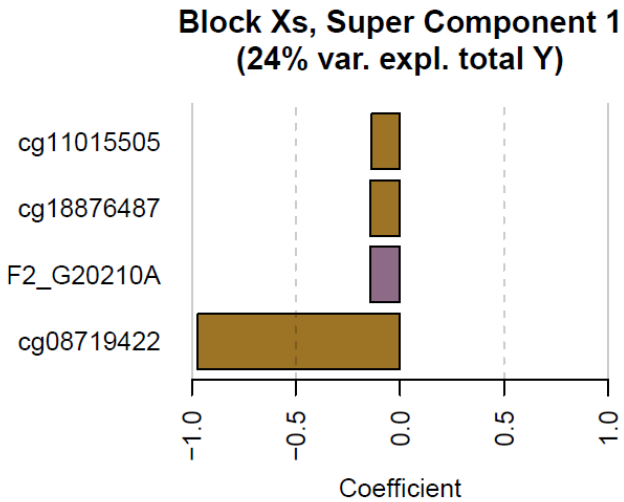
40-folds cross-validation, minimizing MSEP

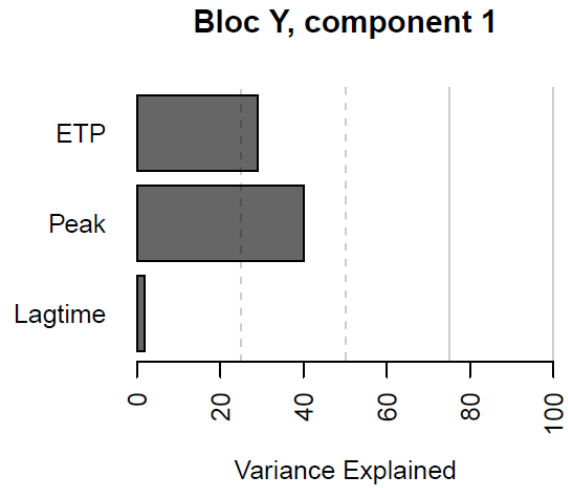


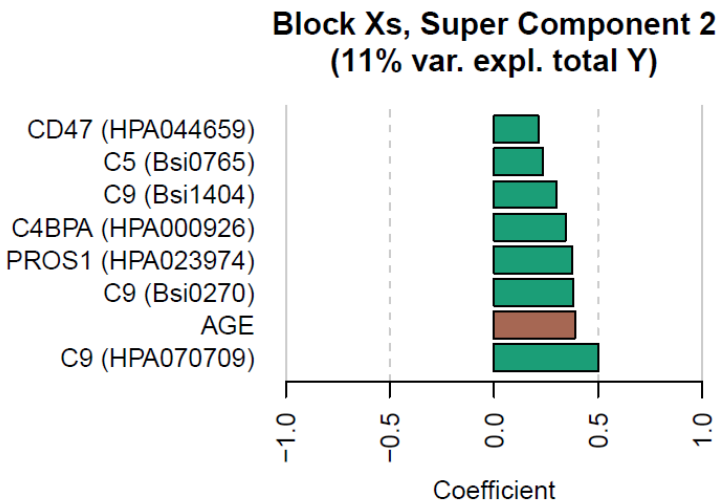
Optimal model

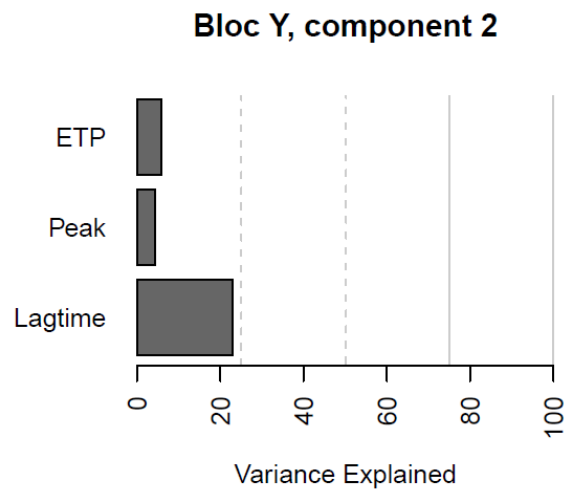
$L_0 = 12,\ R = 2.$

Venous thrombosis data set









Venous thrombosis data set

Observation

The most explained trained information is not the best predicted.

Interpretation

Over-fitting due to correlation structure deformation.

Solution

Use multiple imputation.
But how to adapt ddsPLS to multiple imputation framework ?

Une introduction à l'apprentissage statistique
Décomposition du risque
Compromis biais-variance
Le problème linéaire
Solutions envisagées
La régularisation
My thesis
ddsPLS, complete data
Statistical model and estimators
Application to a real data set
Koh-Lanta, missing values per block
Algorithm
Simulations
Application to real data sets
Ebola data set
Venous thrombosis data set
Packages
Conclusions and perspectives

Packages : ddsPLS for R and py_ddspls for Python

In Python : on GitHub (under-development) and PyPi (stable)

- Build models function,
- cross-validation function,
- parallelized functions.

In R : on GitHub (under-development) and CRAN (stable)

Two functions.

Function	Why ?	Methods
`mddsPLS`	Build models	`summary, plot, predict`
`perf_mddsPLS`	Perform cross-validation	`summary, plot`

C++ and parallelized functions.

Packages : ddsPLS for R and py_ddspls for Python

Various visualizations

Classical data set, $n = 64$, $q = 10$, $p = 3116$.

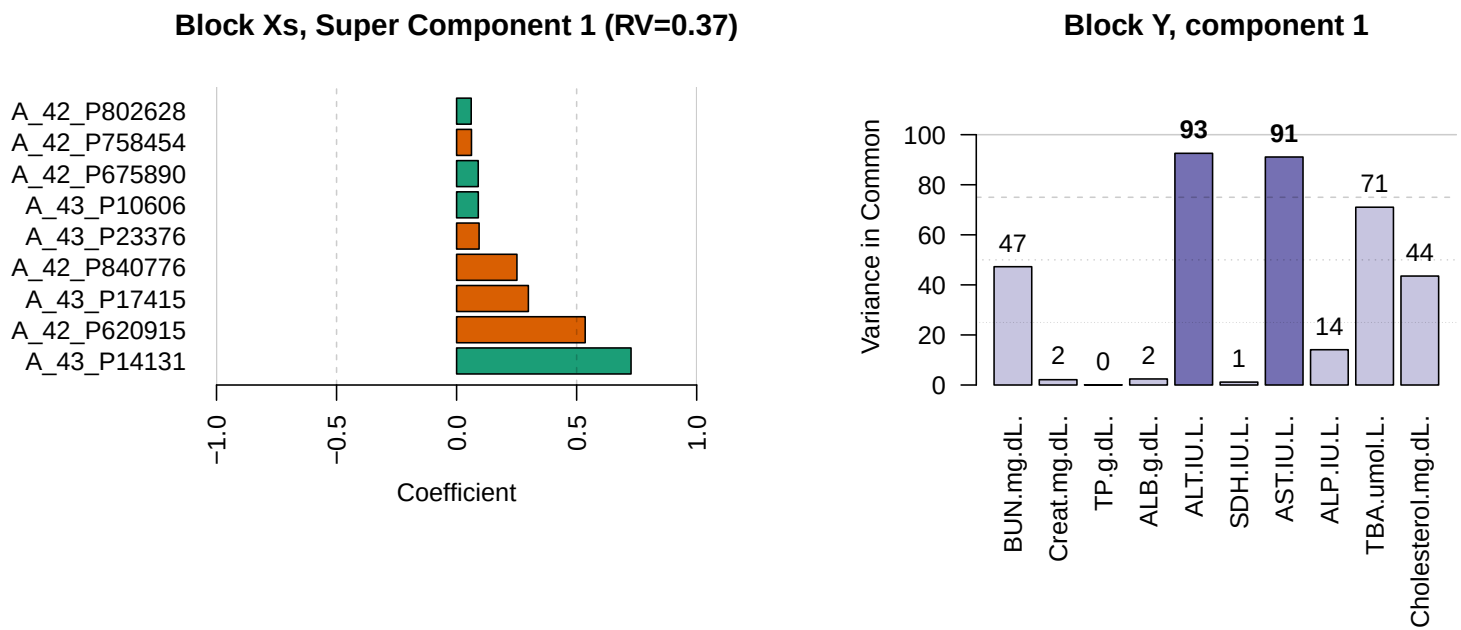
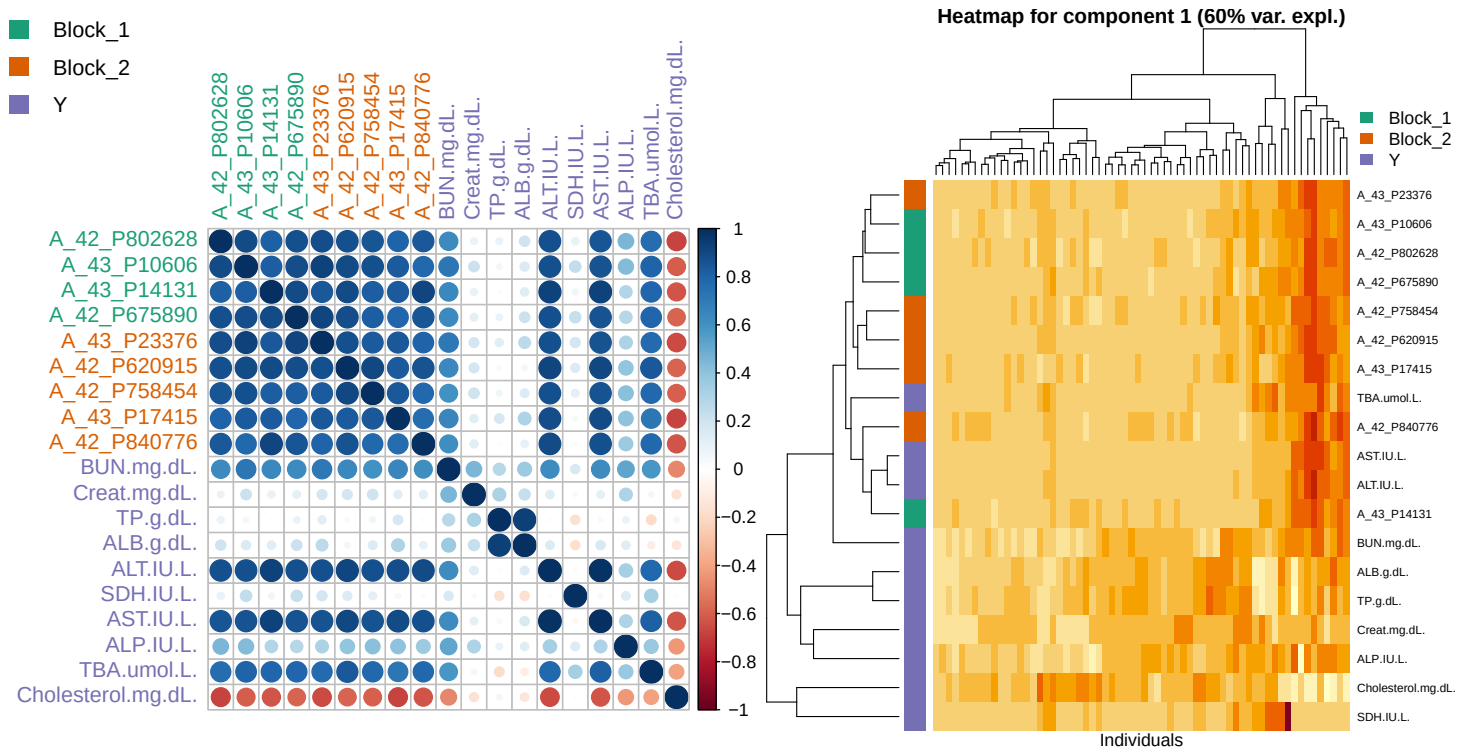


FIGURE – Super-weights, super-components, explained variances.

Packages : ddsPLS for R and py_ddspls for Python

Various visualizations

Classical data set, $n = 64$, $q = 10$, $p = 3116$.



(a) Correlograms.

(b) Heatmaps.

Une introduction à l'apprentissage statistique
- Décomposition du risque
- Compromis biais-variance
- Le problème linéaire
- Solutions envisagées
- La régularisation

My thesis

ddsPLS, complete data
- Statistical model and estimators
- Application to a real data set

Koh-Lanta, missing values per block
- Algorithm
- Simulations
- Application to real data sets
  - Ebola data set
  - Venous thrombosis data set

Packages

Conclusions and perspectives

Conclusions

Methodological results summary

- Multi block,
- missing samples,
- interpretable models,
- variable selection,
- $n << p$ adaptation,
- qualitative with 2 levels,
- regression or classification problems.

Publications

- Methodological article under submition,
- R-journal article submitted,
- real data set applications, three articles. published.

Perspectives

- Adaptation to multiple imputation,
- missing values in **Y**,
- missing values in **X** but not per block,
- investigate bias of the Ridge version,
- Categorical variables in **X** with more than 2 levels.

Thanks

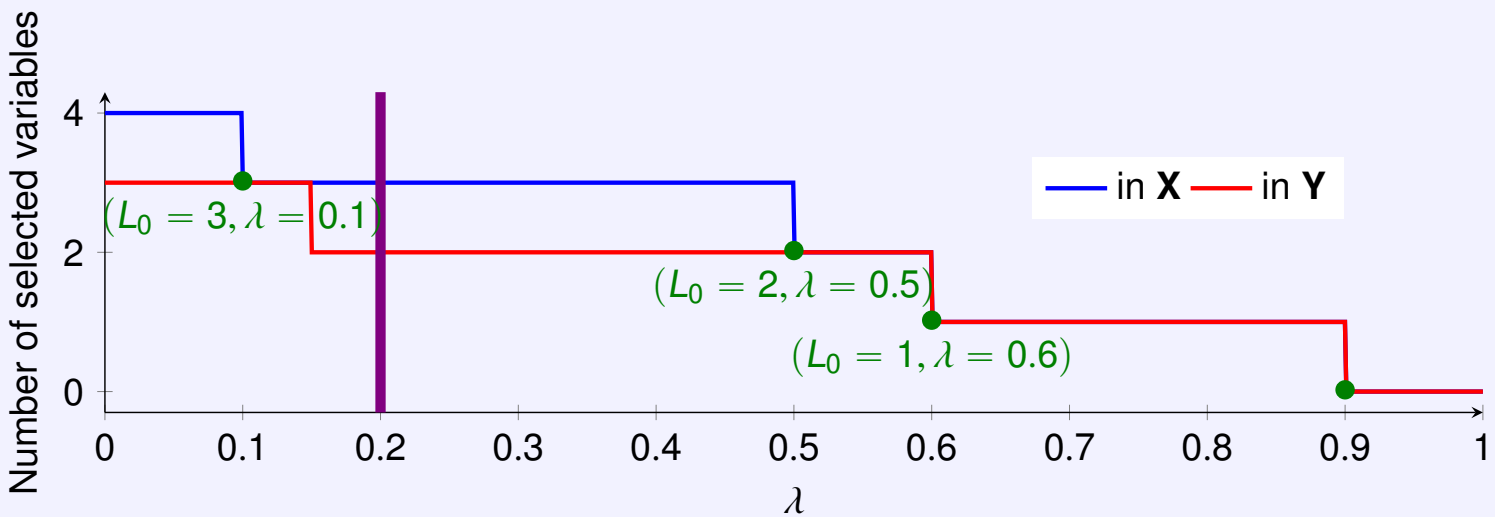
Digital Public Health Graduate Program
université de BORDEAUX
Inria
Inserm
École doctorale Sociétés, politique, santé publique
université de BORDEAUX

Another parametrization

λ and the maximum number of variables in the model

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 & 0.1 \\ 0.5 & -0.2 & 0.01 \\ -0.6 & 0.1 & 0.15 \\ -0.1 & 0.03 & 0.02 \end{bmatrix}}_{\frac{\mathbf{X}^T\mathbf{Y}}{n-1}} \quad \begin{matrix} \Longrightarrow \\ \lambda = 0.2 \end{matrix} \quad \underbrace{\begin{bmatrix} \cdot & 0.7 & \cdot \\ 0.3 & \cdot & \cdot \\ -0.4 & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}_{S_\lambda\left(\frac{\mathbf{X}^T\mathbf{Y}}{n-1}\right)}$$

Definition of L_0 :



Another parametrization

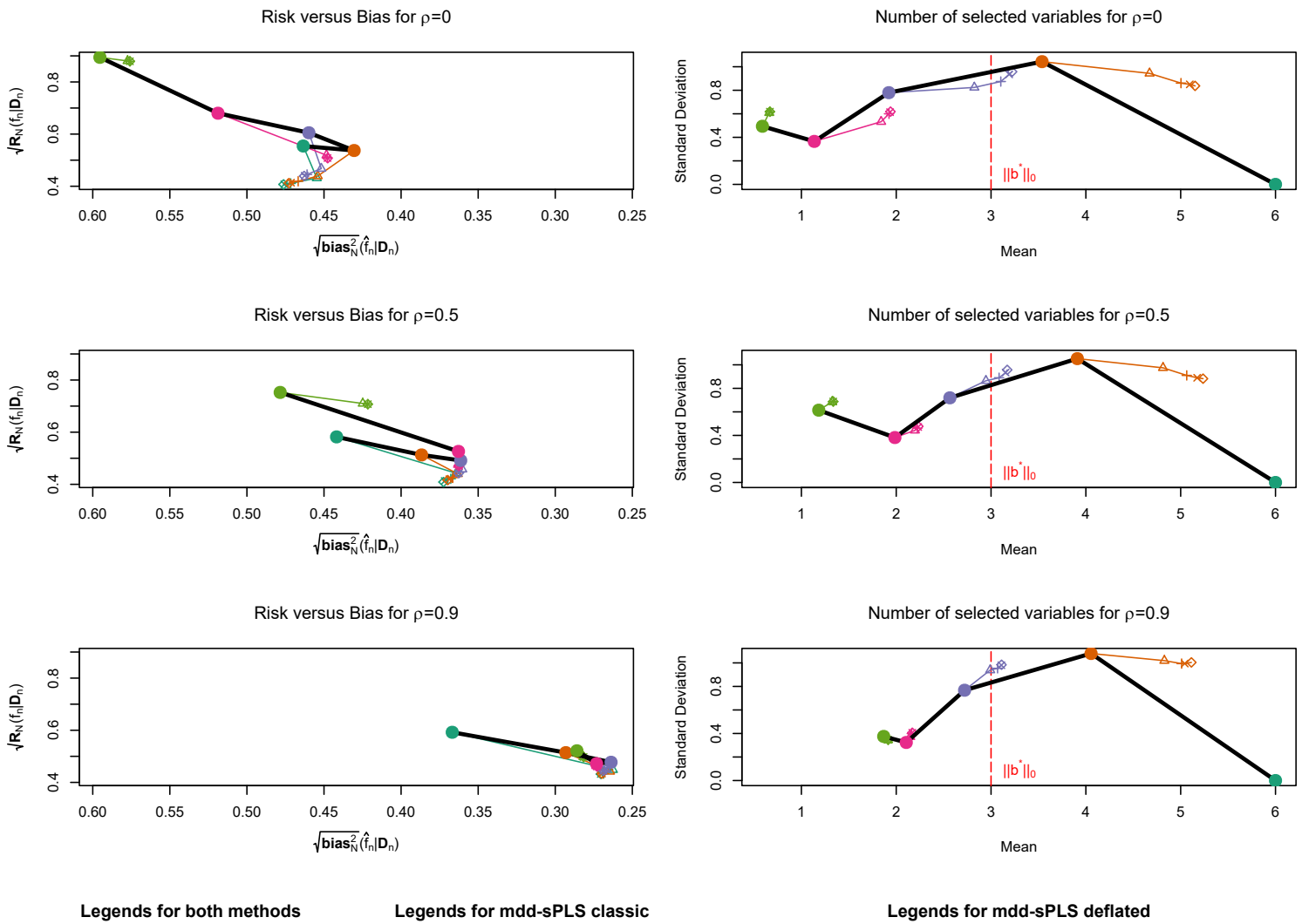
λ and the maximum number of variables in the model

$$\underbrace{\begin{bmatrix} 0.15 & 0.9 & 0.1 \\ 0.5 & -0.2 & 0.01 \\ -0.6 & 0.1 & 0.15 \\ -0.1 & 0.03 & 0.02 \end{bmatrix}}_{\frac{\mathbf{X}^T\mathbf{Y}}{n-1}} \quad \begin{matrix} \Longrightarrow \\ \lambda = 0.2 \end{matrix} \quad \underbrace{\begin{bmatrix} \cdot & 0.7 & \cdot \\ 0.3 & \cdot & \cdot \\ -0.4 & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}}_{S_\lambda\left(\frac{\mathbf{X}^T\mathbf{Y}}{n-1}\right)}$$

Definition of L_0 :

Hypothesis : That parameterization is less sensible to low sample size drawbacks, outliers mainly.
Interest for experts : Closer to what experts seek : a number of predictors.

Deflated ddsPLS simulations



References I


Yash Deshpande et Andrea Montanari. « Sparse PCA via Covariance Thresholding ». In : *Journal of Machine Learning Research* 17.141 (2016), p. 1-41. URL : http://jmlr.org/papers/v17/15-160.html.


Stuart Geman, Elie Bienenstock et René Doursat. « Neural networks and the bias/variance dilemma ». In : *Neural computation* 4.1 (1992), p. 1-58.


Trevor Hastie et al. « Matrix completion and low-rank svd via fast alternating least squares ». In : *Journal of Machine Learning Research* 16.1 (2015), p. 3367-3402. URL : http://jmlr.org/papers/volume16/hastie15a/hastie15a.pdf.


François Husson et Julie Josse. « Handling missing values in multiple factor analysis ». In : *Food quality and preference* 30.2 (2013), p. 77-85.


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).


Hadrien Lorenzo et al. « An Original Methodology for the Selection of Biomarkers of Tenderness in Five Different Muscles ». In : *Foods* 8.6 (2019), p. 206.

References II

Anne Rechtien 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 (sept. 2017), p. 2251-2261. ISSN : 2211-1247. DOI : 10.1016/j.celrep.2017.08.023. URL : http://dx.doi.org/10.1016/j.celrep.2017.08.023.

Vladimir Vapnik. « Principles of risk minimization for learning theory ». In : *Advances in neural information processing systems* (1992), p. 831-838.

Vladimir N Vapnik et A Ya Chervonenkis. « On the uniform convergence of relative frequencies of events to their probabilities ». In : *Measures of complexity*. Springer, 2015, p. 11-30.