High-dimensional multi-block analysis of factors associated with thrombin generation potential

Hadrien Lorenzo, Misbah Razzaq, Jacob Odeberg,
Pierre-Emmanuel Morange, Jérôme Saracco,
David-Alexandre Trégouët, Rodolphe Thiébaut

Bordeaux, France
hadrien.lorenzo@u-bordeaux.fr

June 6, 2019

Venous Thrombosis and Thrombin Generation Potential

Venous Thrombosis, a complex disease

Characterized by

- Formation of a blood clot in a vein,
- Clot can break free $\rightsquigarrow$ Lung $\longrightarrow$ Pulmonary embolism.

$\Longrightarrow$ 3^{rd} major cause of cardiovascular disease.

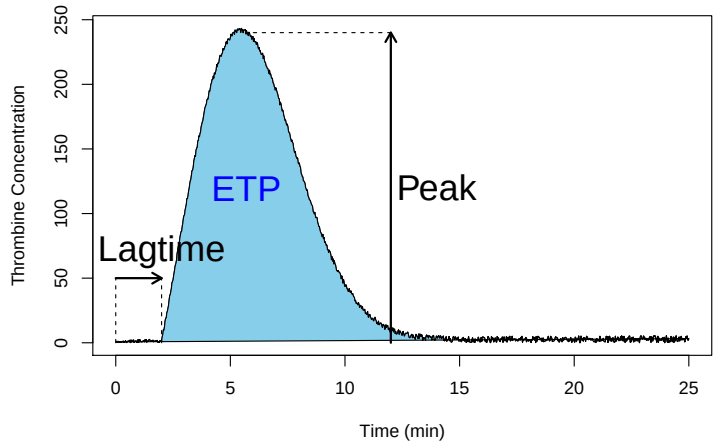
Thrombin, a key molecule in the coagulation cascade

High level of thrombin $\Longrightarrow$ Risk factor for Venous Thrombosis.

Dynamics of the Thrombin Generation

3 main biomarkers

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



Main known factors influencing the Thrombin Generation

- Age, Sex, BMI (*Body Mass Index*).
- Mutation *F2 G20210A* linked to Peak and ETP,
  [Rocanin-Arjo et al., 2014].

Multi-omics data sets

696 patients with Venous Thrombosis from the MARTHA cohort:

- **3** main biomarkers of the Thrombin Generation.
- **384** plasma biomarker proteins.
- $\approx$ **3000** whole blood DNA CpG sites (DNA methylation data)
- *F2 G20210A* mutation.
- **3** main natural coagulation inhibitors:
  Protein S, Protein C and Antithrombin.
- Phenotypes: Age, Sex, BMI.

Missing values challenge

$\approx$ 32% of the DNA methylation: sub-study randomly sampled,

$$\Longrightarrow \approx 68\% \text{ of missing values.}$$

Objectives

Find a model that:

- Identify the most relevant biomarkers linked to the 3 main thrombin generation biomarkers.
- Takes into account samples with missing values.

Additional mathematical challenges

- Number of participants lower than number of features.
  $\Longrightarrow$ High dimensional setting
- Different data types for each individual
  $\Longrightarrow$ Multi-block heterogeneous data

Data-Driven Sparse Partial Least Square (ddsPLS) [Lorenzo, Saracco, and Thiébaut, 2019]

Idea

Given a covariate matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ that should predict a response matrix $\mathbf{Y} \in \mathbb{R}^{n \times q}$, n the number of individuals.
Describe covariance structure $< \mathbf{Y}, \mathbf{X} >$ under some thresholding asumption.
Chosen solution, close to [Deshpande and Montanari, 2016]:

Do the **SVD** decomposition of the **soft-thresholded covariance matrix** over **R** components.

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

Soft-Thresholding, $\forall \lambda \in [0,1]$: $S_\lambda(t) = \begin{cases} 0 & \text{si } |t| < \lambda \\ t - \lambda & \text{si } t \geqslant \lambda \\ t + \lambda & \text{si } t \leqslant -\lambda \end{cases}$

Multi-Block case with missing values

Considering T blocks in the covariate part. 2 steps alternation : Data structure decomposition and Missing values estimation.

Data structure decomposition : 3 steps solution

1. Per block:
   $\forall t=1..T, \mathbf{U}_t = (u_t^{(1)}, .., u_t^{(R)}) = \arg\max_{\mathbf{u}^T\mathbf{u}=\mathbb{I}_R} ||S_\lambda\left(\frac{\mathbf{Y}^T\mathbf{X}_t}{n-1}\right)\mathbf{u}||_F^2,$
2. Aggregate the blocks:
   $\mathbf{M}_t = S_\lambda(\frac{\mathbf{Y}^T\mathbf{X}_t}{n-1}),\ \mathbf{Z} = [\mathbf{M}_1\mathbf{U}_1, \cdots, \mathbf{M}_T\mathbf{U}_T],$
   $\underline{\beta} = [\underline{\beta}_1^T, \cdots, \underline{\beta}_T^T]^T = \arg\max_{\underline{\beta}^T\underline{\beta}=\mathbb{I}_R} ||\mathbf{Z}\underline{\beta}||_F^2 \in \mathbb{R}^{RT\times R}$
3. The regression model: $\mathbf{Y} \approx \sum_{t=1}^T \mathbf{X}_t\mathbf{B}_t$
   $\mathbf{B}_t = \mathbf{U}_t^\star\mathbf{B}_0\mathbf{V}^{\star T},\ \mathbf{B}_0 = (\mathbf{T}^{\star T}\mathbf{T}^\star)^+\mathbf{T}^{\star T}\mathbf{S}^\star,\ \mathbf{U}_t^\star = \mathbf{U}_t\underline{\beta}_t,$
   $\mathbf{V}^\star = norm_2(\mathbf{Z}\underline{\beta}),\ \mathbf{T}^\star = \sum_{t=1}^T \mathbf{X}_t\mathbf{U}_t^\star,\ \mathbf{S}^\star = \mathbf{Y}\mathbf{V}^\star.$

Missing values estimation

Thanks to what follows...

Missing values estimation

$$\mathcal{X} = (\mathbf{X}_1, .., \mathbf{X}_T) = \left(\begin{array}{c} \mathbf{X}_1^{(train)} \\ \mathbf{X}_1^{(test)} \end{array} \; ... \; \begin{array}{c} \mathbf{X}_T^{(train)} \\ \mathbf{X}_T^{(test)} \end{array} \right) \qquad \mathbf{Y} = \begin{array}{c} \mathbf{Y}^{(train)} \\ \mathbf{Y}^{(test)} \end{array}$$

Two different algorithms depending on the step: **train** or **test**:

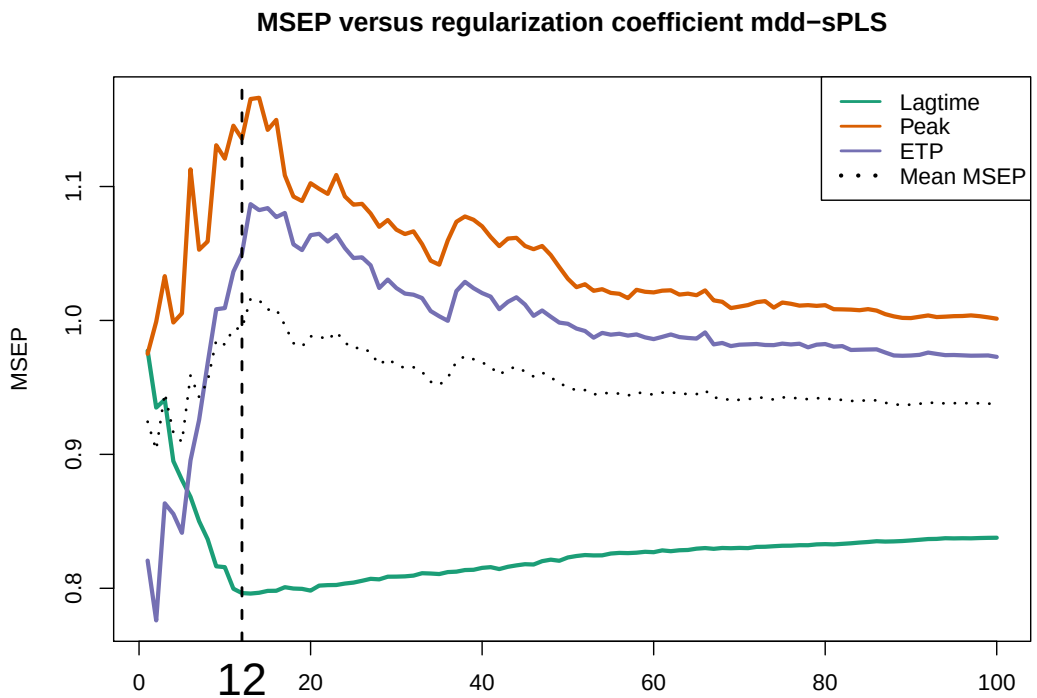
- The **train** data set imputation **uses** $\mathbf{Y}^{(train)}$.
- The **test** data set imputation **DOES NOT use** $\mathbf{Y}^{(test)}$.

Application to **MARTHA** cohort prediction of thrombin

Choice of the 2 parameters

- L_0, the maximum number of selected covariables,
- R, the number of components:
  up to 3 (number of thrombin biomarkers).

Fixed thanks to 40-folds cross-validation, minimizing MSEP.



Optimal model

$L_0 = 12,\ R = 2.$

The identified dd-sPLS model

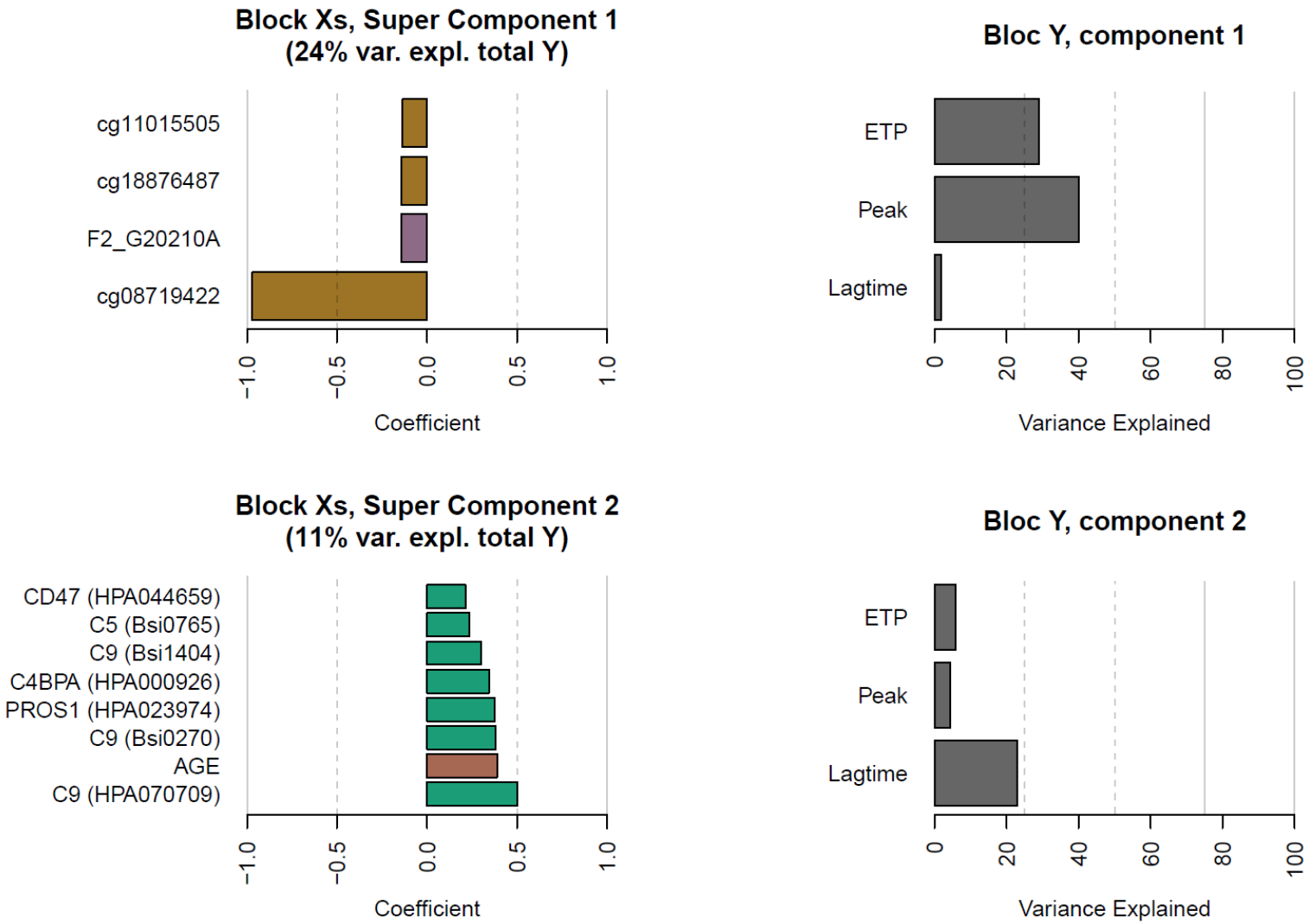


Figure: Scaled super-weights per super-component and variance explained per response variable per component.

Preliminary replication study

- 133 independant venous thrombosis patients,
- No DNA methylation measurements

$$\Longrightarrow \quad 1^{st} \text{ component not computable}$$

- Projection on the 2^{nd} component:

$$r = 0.16,\ p = 0.069,$$

while r = 0.23 for the training data set.

Conclusion and future works

- Successful application of a new multi-omics method,
- Clinically: a signature composed of 7 proteins explain 20% of the *LagTime*.
  $\Longrightarrow$ Further clinical investigations on healthy patients.
- Extraction of information from the methylation data set while 70% of missing values.
- Packages available on the **CRAN**, **PyPi** and **GitHub** (to be preferred for now),

```
> devtools::install_github("hlorenzo/ddsPLS")
```

Thanks!

References


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


Trevor Hastie et al. “Matrix completion and low-rank SVD via fast alternating least squares”. In: *The Journal of Machine Learning Research* 16.1 (2015), pp. 3367–3402.


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


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, Jérôme Saracco, and Rodolphe Thiébaut. “Supervised Learning for Multi-Block Incomplete Data”. In: *arXiv preprint arXiv:1901.04380* (2019).

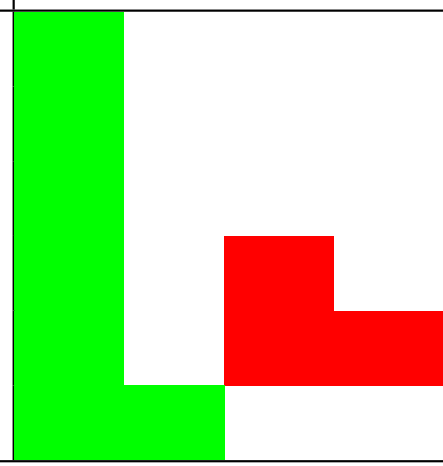

Kristiaan Pelckmans et al. “Handling missing values in support vector machine classifiers”. In: *Neural Networks* 18.5-6 (2005), pp. 684–692.


Ares Rocanin-Arjo et al. “A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential”. In: *Blood* 123.5 (2014), pp. 777–785.


Daniel J. Stekhoven and Peter Buehlmann. “MissForest - non-parametric missing value imputation for mixed-type data”. In: *Bioinformatics* 28.1 (2012), pp. 112–118.

Available methods

Method	Imputation	Multi-block	Time-consuming	Sample size requirement	Variable Selection	Supervised
Mean						
Nipals [Lê Cao et al., 2008]						
softImpute [Hastie et al., 2015]						
missForest [Stekhoven and Buehlmann, 2012]						
SVM [Pelckmans et al., 2005]						
imputeMFA [Josse and Husson, 2016]						



Simulation results

Prediction error against proportion of missing samples.

10 blocks, 100 individuals, 160 variables.

3 components but only one correlated with the univariate response.